# A PERFORMANCE SURVEY OF TEXT-BASED SENTIMENT ANALYSIS METHODS FOR AUTOMATING USABILITY EVALUATIONS

A Thesis

presented to

the Faculty of California Polytechnic State University,

San Luis Obispo

In Partial Fulfillment

of the Requirements for the Degree

Master of Science in Computer Science

by

Kelsi Rado Van Damme

June 2021

## COMMITTEE MEMBERSHIP

| | |
|---|---|
| TITLE: | A Performance Survey of Text-Based Sentiment Analysis Methods for Automating Usability Evaluations |
| AUTHOR: | Kelsi R. Van Damme |
| DATE SUBMITTED: | June 2021 |
| COMMITTEE CHAIR: | Franz Kurfess, Ph.D.<br>Professor of Computer Science |
| COMMITTEE MEMBER: | Dongfeng Fang, Ph.D.<br>Professor of Computer Science |
| COMMITTEE MEMBER: | Sandrine Fisher, Ph.D<br>Human-Computer Interaction Researcher |

## Abstract

### A Performance Survey of Text-Based Sentiment Analysis Methods for Automating Usability Evaluations

Kelsi Rado Van Damme

Usability testing, or user experience (UX) testing, is increasingly recognized as an important part of the user interface design process. However, evaluating usability tests can be expensive in terms of time and resources and can lack consistency between human evaluators. This makes automation an appealing expansion or alternative to conventional usability techniques.

Early usability automation focused on evaluating human behavior through quantitative metrics but the explosion of opinion mining and sentiment analysis applications in recent decades has led to exciting new possibilities for usability evaluation methods.

This paper presents a survey of modern, open-source sentiment analyzers' usefulness in extracting and correctly identifying moments of semantic significance in the context of recorded mock usability evaluations. Though our results did not find a text-based sentiment analyzer that could correctly parse moments as well as human evaluators, one analyzer was found to be able to parse positive moments found through audio-only cues as well as human evaluators. Further research into adjusting settings on current sentiment analyzers for usability evaluations and using multimodal tools instead of text-based analyzers could produce valuable tools for usability evaluations when used in conjunction with human evaluators.

## Acknowledgements

Thank you to:

- Gavin K. Chao for his continued support and friendship throughout our complementary theses and from my very first day at Cal Poly

- Dr. Franz J. Kurfess for his never ending encouragement and incredible guidance as a mentor, as my advisor, and as my committee chair

- Dr. Dongfeng Fang for giving me a sense of direction early in my Master's career and for her advice as a committee member

- Dr. Sandrine Fischer for sharing her insight into the field and for her advice as a committee member

- Erin Sheets for allowing us to gather data in her courses and for her recommendations in our usability tests

- Andrew Barney for providing me with a loving foundation on which I could grow and thrive in my studies

## TABLE OF CONTENTS

Page

CHAPTER

# LIST OF TABLES

Table Page

# LIST OF FIGURES

Figure Page

# Chapter 1 Introduction

The turn of the century saw the rise of big data and with it, the modernization of sentiment analysis. Today Twitter, Facebook, Reddit, Google, and millions of other spaces hold incredible written collections of opinionated data. Such remarkable volumes of thought and expression have become a wealth of information for opinion mining and text-based sentiment analysis. These automated tools aim to detect and extract subjective information, like opinions and attitudes, from written texts.

Together with big data, text-based sentiment analysis has proliferated into a variety of fields. In politics, sentiment analysis is being used on social media to track candidates' popularity and create voter polls [6]. In market management, companies use sentiment analysis to shape brand image and in public health, sentiment analysis has been used to detect depression in patients [5]. Yet despite its wide-spread use, there remain many unexplored applications of sentiment analysis. Accordingly, this paper seeks to survey one such, scarcely studied avenue: the augmentation and automation of usability evaluations.

The benefits of usability evaluations to end users and companies is widely undisputed. Karat cost-benefit analyses have shown for well over a decade that usability testing provides between a 2:1 and 100:1 savings-to-cost ratio for software development projects [2]. However, user testing methods continue to face the same practical challenges. Moderated methods are expensive in terms of time and resources, and can lack scalability, coverage, and consistency between moderators [1]. Unmoderated tests that use automated evaluations widely solve these issues but lack the ability to collect and resolve qualitative data [1, 10].

One potential solution is the adaptation of sentiment analysis to capture qualitative data in remote unmoderated usability tests. The ability to automate affectual data collection, especially in conjunction with existing methods of automatic quantitative data collection, could lead to a wide variety of tools that enhance current usability evaluations.

To help understand the current value of text-based sentiment analysis for improving usability evaluations, this paper conducts a performance survey of four representative techniques: AFINN [14], SentiStrength [16], Umigon [18], and VADER [20]. The evaluation focuses on each analyzer's ability to detect and classify sentence polarity in the context of usability evaluations. This is achieved through comparison against a created and human-labeled dataset of transcribed mock usability tests.

# Chapter 2 Background & Related Work

The following section provides important definitions and justifications for the focus of our performance survey.

## 2.1 Accomodations for COVID-19

All of the experiments and evaluations in this paper were conducted under the restrictions of COVID-19. As such, some decisions were made to adapt to the circumstances.

For our data gathering experiments, all participants were recruited from two, virtually taught, courses of CSC 486: Human Computer Interaction. Due to the remote nature of our interactions, we chose to use two websites as the targets of our usability tests. Similarly, uniform recordings of the tests could not be taken remotely without large overhead. Instead, participants were asked to record themselves on their regular home setup through their Cal Poly Zoom accounts, with specific formatting instructions. Though, there was still some variation in the quality of these recordings, for the purposes of our examination methods, we found this to be negligible.

## 2.2 Designing Mock Usability Tests

The user testing method used to generate our testing dataset followed a think aloud protocol and standard remote, unmoderated usability testing techniques.

Think aloud protocols ask participants to verbalize their thoughts and opinions as they move through the product interface. These protocols are commonplace in most user testing methods and are seen as one of the most effective ways to gather reasoning and affectual data [7]. This kind of verbal data also allowed us to later transcribe participants' speech into text as a suitable input medium for testing sentiment analysis tools.

As previously mentioned, we also chose remote usability testing to accommodate for socially distanced participants. Remote methods primarily fall into two categories: moderated testing and unmoderated testing. In moderated tests, an evaluator observes participants and facilitates the test in real time, while unmoderated tests allow users to complete written tasks on their own, and any data collected during the test is reviewed at a later time [7]. Though moderated methods are more common, studies have shown that moderated and unmoderated users make largely the same verbalizations when thinking aloud [8]. As the value of verbalizations is equivalent between the two techniques, we chose to use unmoderated methods because they better align with our pursuit of automation and for ease of implementation.

With these considerations in mind, remote unmoderated usability tests that followed a think aloud protocol, were used to generate our final testing dataset. Greater details on the layout of our mock usability tests are shown in Table 4.1 under Chapter 4: Methods.

### 2.3 Focus on Sentence-Level Polarity Detection and Classification

Methods of sentiment analysis can often be applied to a variety of tasks and can perform a variety of functions. Some analyzers attempt to recognize specific emotions, like happy or sad, while others will aim to recognize the subject that a particular sentiment is directed towards. Often, these various functions are not directly comparable so, for the purposes of this evaluation, we restricted our focus to evaluating efforts related to detecting the polarity (positivity or negativity) of transcribed usability tests. Polarity detection is a common functionality among sentiment analyzers and can provide valuable affectual information.

During performance evaluations, detection and classification was run on a sentence-by-sentence level. This granularity was chosen based on observations of the labels in our dataset. Generally, moments flagged by participants were short, relating to things like a small laugh, or a sentence. Flagged moments longer than a sentence were rare and, as such, we found sentence-level polarity detection to best reflect human-identified moments.

# Chapter 3 Sentiment Analysis Methods

This section provides a brief description of the four text-based sentiment analysis methods investigated in this paper. All tools reviewed were available for download off the web and were not changed or adjusted in any way other than to analyze input at a sentence-by-sentence level.

All four text-based sentiment analyzers were chosen from methods found to be high performing by the benchmark study, *SentiBench : a benchmark comparison of state-of-the-practice sentiment analysis methods* [3]. The Sentibench study evaluated twenty-four popular text-based sentiment analyzers based on a benchmark of eighteen gold-standard datasets. Datasets used in the article covered messages posted on social networks, movie and product reviews, and opinions and comments in news articles.

## 3.1 Selection Criteria

In addition to using methods reviewed in the Sentibench study, three other requirements were created to better select for methods that aligned with the goals of this study:

1. A method must have scored a mean rank above tier 7 in either the 3-class or 2-class evaluations of the Sentibench study.
2. The method must detect and categorize sentence polarity using three distinct classes: positive, neutral, and negative.
3. The method must be free and readily accessible to the public.

Due to time constraints, not every sentiment analysis method that we would've liked to test could be tested. Therefore to help choose a well-suited testing set, we considered techniques that had already shown a high level of performance first. Consequently, methods that did not score a mean rank above 7th in the Sentibench study were left out. This does not however mean that methods excluded for this reason would not have performed well under the context of usability testing, and they may be of interest in future research.

Requirement two was intended to maintain congruent performance evaluations. Polarity detection is a common functionality among sentiment analyzers, however, there is some variability in its implementation. Typically classification of polarity is implemented using either a two-class system (positive, negative) or a three-class system (positive, neutral, negative). Because human-labeling of the dataset was better suited to a three-class system, we required the output of all tested sentiment analyzers to match a three-class system. This condition allowed for direct comparisons between the human-labeled dataset and the tested analyzers during the performance evaluation.

Requirement three was implemented primarily for ease of access. In most cases, the authors of closed source methods could not be reached, creating barriers of availability, while fiscal restraints prevented our team from access to paid methods. Table 3.1 below provides an overview of each analyzer used in this study, including a brief description and general statistics about each tool.

**Table 3.1 Overview of the Text-based Sentiment Analyzers Reviewed in this Literature**

| Name | Description | Output | Lexicon Size | Machine Learning |
|---|---|---|---|---|
| AFINN [12, 14] | Builds a Twitter based sentiment Lexicon including Internet slangs and obscene words. AFINN can be considered as an expansion of ANEW [13], a dictionary created to provide emotional ratings for English words. ANEW dictionary rates words in terms of pleasure, arousal and dominance. | Provides polarity score for lexicons (−5 to 5) | 2,477 | -- |
| SentiStrength [15, 16] | Builds a lexicon dictionary annotated by humans and improved with the use of Machine Learning. | (−5, 2), −1, 1, (2, 5) | 2,698 | ✓ |
| Umigon [17, 18] | Disambiguates tweets using lexicon with heuristics to detect negations plus elongated words and hashtags evaluation | Negative, Neutral, Positive | 1,053 | -- |

| | | | | |
|---|---|---|---|---|
| VADER [19, 20] | A sentiment analysis method developed for Twitter and social media contexts. VADER was created from a generalizable, valence-based, human-curated gold standard sentiment lexicon. | [< –0.05),<br>(–0.05, ...,0.05),<br>(> 0.05] | 7,517 | -- |

# Chapter 4 Dataset Collection

To evaluate the sentiment analysis methods outlined in Section 3, a small-scale experiment was performed to gather a testing dataset. The experiment was conducted in partnership with Gavin Chao who also uses the collected dataset in his related paper, *Applying Facial Emotion Recognition to Usability Evaluations to Reduce Analysis Time* [4]. Section 4 below outlines our data gathering experiment and the following section outlines the performance evaluation methods used for testing each sentiment analysis tool.

## 4.1 Gathering Data

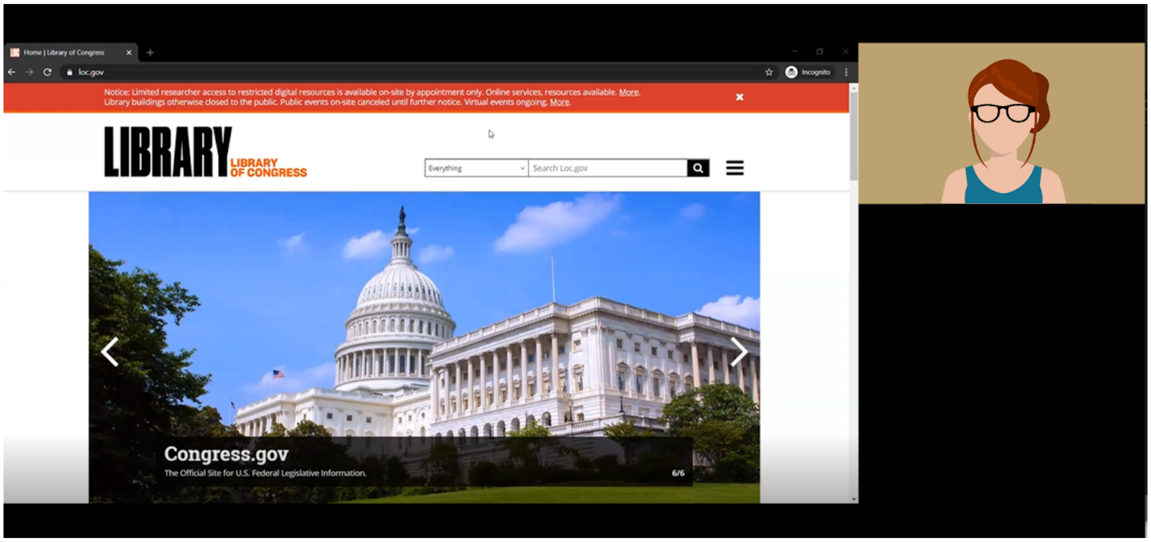
Datasets of recorded usability tests are scarce. Publicly releasing usability testing of a product under development provides little benefit to companies, and may clash with company privacy policies or increase company liability related to participant consent. These limitations give companies little reason to post usability recordings and there exists a lack of open-source usability testing datasets.

To fill this gap, mock usability tests were conducted on students in two courses of CSC 486: Human-Computer Interaction. Students in each course were given the option to voluntarily participate in the experiment for course credit or complete an alternative assignment given by the instructor. Students were encouraged to be as candid as possible during their participation and it was emphasized that students would not be graded on their performance during the experiment. Between the two classes, 39 students out of 43 chose to participate. Of the 39 submissions received, 35 recordings were deemed valid for use in the final dataset.

The experiment in each course was conducted using the same two-stage method over a two week period. In the first stage students record themselves completing an unmoderated usability test to add to the dataset. In the second stage students are given recordings from the dataset, other than their own, to tag for moments of semantic significance that could later be compared against the output of an automated evaluator. For the purposes of this experiment, semantic significance was defined as a display, by the subject, of evaluative judgment, such as positive or negative, or an emotional or affectual attitude such as frustration, joy, anger, sadness, excitement, and so on [5].

### 4.2 Adding Usability Tests to the Dataset

At the beginning of the experiment, participants in each course were randomly divided into two groups. Students in Group 1 were asked to record themselves completing a usability test for the website www.loc.gov, while students in Group 2 were asked to record themselves completing a different usability test for the website www.ca.gov. As both courses were taught remotely, students were asked to record themselves on their regular home setup through their Cal Poly Zoom accounts. An example of the required recording format is shown below in Figure 4.1.



Notice: Limited researcher access to restricted digital resources is available on-site by appointment only. Online services, resources available. More.
Library buildings otherwise closed to the public. Public events on-site canceled until further notice. Virtual events ongoing. More.

Home | Library of Congress
loc.gov
Incognito
LIBRARY
LIBRARY OF CONGRESS
Everything
Search Loc.gov
Congress.gov
The Official Site for U.S. Federal Legislative Information.
6/6

Figure 4.1 : Example Recording Layout of A Mock Usability Test

Both usability tests were designed as remote, unmoderated usability tests to accommodate remote learning and to better align with the goals of our evaluation. During the test, participants were asked to follow a think aloud protocol, in which they verbalize their thought processes and opinions during the test. These verbalizations allowed us to later transcribe participants' speech into text as a suitable input medium for testing sentiment analysis tools. Each test consisted of 5 tasks intended to take a total of 15 minutes to complete. A brief outline of the websites and tasks used in each usability test are shown in table 4.1.

**Table 4.1 : Overview of the Tasks in each Mock Usability Test**

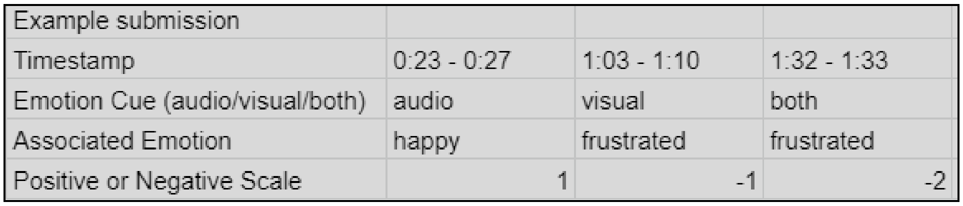
| Task Number | **Group 1**<br>Website: www.loc.gov | **Group 2**<br>Website: www.ca.gov |
|---|---|---|
| 1 | Find any text that contains George Washington's Farewell Address and be able to read it from the screen. **(DO NOT READ THE FAREWELL ADDRESS! Just get to a page where you can read it.)** | Find the number of fires since the start of this year. |
| 2 | Find who has access to the physical library and how to gain access to the physical library. | Find the dataset for COVID-19 Tests and look for the tests completed as of your current date. |
| 3 | Find the steps to register for a copyright. | Find the COVID-19 Information from the California Department of Aging. Get to the page for COVID-19 Resources for American Sign Language. |
| 4 | Buy a framed print of the Gettysburg Address. **(Do not actually buy the item, just get to the screen where you enter your information and stop there)** | Find the official voter information guide for California from the California Secretary of State. |
| 5 | Find how to get a Reader Registration Card. | Find what a Blue Alert is on the CHP section of the website. |

### 4.3 Tagging the Dataset

Stage 2 of the experiment began at the start of the second week. Students were randomly assigned 3 recordings from the opposite testing group to tag for moments of semantic significance. Previous studies have shown that non-expert labeling may be as effective as labels created by experts for affect recognition [9], and having students label multiple recordings allowed for multiple human assessments of each usability test. For each moment identified in a recording, students were asked to document the following:

1. The time range in the video where the moment occurred
2. Whether the moment was identified through a visual cue, audio cue, or both
3. The emotion most closely associated with the moment
4. The valence of the moment on a whole-number scale from -2, being the most negative to +2 being the most positive

An abbreviated example of tags written for a recording is shown in Figure 4.2. Unlabeled moments were assumed to be neutral with a valence of 0 and no associated emotion. As the collected data set was intended for use in other studies, more information was tagged than is used in the performance evaluations of this study.



| | | | |
|---|---|---|---|
| Example submission | | | |
| Timestamp | 0:23 - 0:27 | 1:03 - 1:10 | 1:32 - 1:33 |
| Emotion Cue (audio/visual/both) | audio | visual | both |
| Associated Emotion | happy | frustrated | frustrated |
| Positive or Negative Scale | 1 | -1 | -2 |

**Figure 4.2 : An Example Label for a Usability Test Recording**

Once all participants turned in their labels, the raw dataset was collated, reviewed, and cleansed. Recordings with errors like incorrect formatting, corrupted files, or where the participant forgot to record themselves the first time they took the test, were removed from the dataset. Incomplete sets of tags and improperly formatted sets of tags were also removed from the dataset. Of the 39 tagged recordings received, 35 were deemed usable in the final mock usability testing dataset.

Recordings that were kept had their labels reviewed and systematically combined into a single timeline for use in the performance evaluation. In cases where multiple students marked the same moment as semantically significant, tags were merged based on a simple majority with the time range extended from the earliest marked time to the latest marked of the predominant group of tags. If a simple majority could not be reached, tie-breakers were resolved based on the discretion of the reviewer.

# Chapter 5 Performance Evaluation Metrics

To evaluate a sentiment analyzer's ability to extract and correctly identify moments of semantic significance, recordings of user tests had to first be transformed into a suitable input medium. Each of the 35 tests in our dataset were manually transcribed into separate text files and line-separated by sentence. Each sentence was then run through each analyzer and given a polarity score of positive, negative, or neutral. This paper uses traditional accuracy, precision, recall, and F1 metrics to compare the predicted polarities against the actual human-labeled polarities from the dataset. Table 5.1 below represents the confusion matrix for the analysis.

**Table 5.1 : Confusion Matrix for Three Classification Output of Polarity**

| | | **Predicted (Y)** | | |
|---|---|---|---|---|
| | | positive | neutral | negative |
| **Actual (X)** | positive | a | b | c |
| | neutral | d | e | f |
| | negative | g | h | i |

Each letter in the table represents the number of instances that actually have a polarity of X but were predicted as Y by the sentiment analyzer, where X; Y ∈ { positive; neutral; negative }. Definitions and example equations for each metric used are as follows:

- The **accuracy** of each analyzer is the ratio of correctly predicted sentences to the total number of sentences:

$$Accuracy \ = \ \frac{a+e+i}{a+b+c+d+e+f+g+h+i}$$

- The **precision** of a polarity X is the ratio of sentences correctly predicted as X to the total number of sentences predicted as X:

$$Precision_{positive} = \frac{a}{a+d+g}$$

- The **recall** of a polarity X is the ratio of sentences correctly predicted as X to the actual number of sentences that are X:

$$Recall_{negative} = \frac{i}{g+h+i}$$

- The **F1-score** measures the harmonic mean between both precision and recall:

$$F1\ score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

A total of 3,603 sentences were used for this analysis, with 762 sentences labeled by human evaluators and 2,841 unlabeled sentences. Almost all labeled sentences were labeled as positive or negative with only 2.5% of sentences labeled as neutral by participants. All unlabeled sentences were assumed to be neutral and recognized through both audio and visual cues.

The distribution of the polarity of each sentence used in our analysis is shown in Table 5.2. Notably, neutral moments made up the vast majority of the dataset. These moments constitute almost 80% of the data while positive and negative moments each make up approximately 10% of the data.

**Table 5.2 : The Polarity of Human-Labeled Moments by Sentence**

| positive | neutral | negative | Total |
|---|---|---|---|
| 333 | 2,867 | 403 | 3,603 |

All unlabeled moments were assumed to be found through both visual and audio cues indicating the absence of sentiment. Under this assumption, 79% of the data was observed through both visual and audio cues. However, of the sentences actually labeled by human participants, 9% were found through visual only cues, 38% were found through audio only cues and 53% were found through both. This distribution is shown in Table 5.3.

Table 5.3 : The Form of Expression Used to Recognize Human-Labeled Moments by Sentence

| visual | audio | both | Total |
|---|---|---|---|
| 70 | 289 | 403 | 762 |

This distribution indicates that for our mock usability tests, evaluators utilized audio cues more heavily than visual cues to recognize affectual data, but that both types of cues are significant and necessary to recognize moments of semantic significance.

### 5.1 Study of Human Agreement

To be considered valuable for user experience testing, it is reasonable to expect a sentiment analysis tool to predict the polarity of a moment at least as well as a human evaluator. However, the predictive ability of human evaluators is not perfect. Some studies have shown that when evaluating the polarity of sentences, human analysts tend to agree around 80-85% of the time [25, 26]. Other studies have shown that when evaluating usability tests with the same methodology, independent usability testing teams can have less than a 1% overlap in findings [30, 1].

In our own dataset, we found the average agreement between any two evaluators to be 84% with a weighted kappa of 0.445 . A kappa coefficient measures interobserver reliability while correcting for agreement that may occur through chance [27]. A weighted kappa accounts for the degree of disagreement between two observers. For example, observers who rated the same sentence as positive and neutral would show better agreement than observers who rated the same sentence as positive and negative. According to Landis and Koch, an average weighted kappa of 0.445 shows moderate to low agreement between two evaluators [28]. The statistics and weight map used to measure the average agreement between any two evaluators in this sub-study is shown in Figure 5.1. The methodology used to calculate weighted kappas can be found at real-statisitcs.com [29].

**Figure 5.1 : Study of Human Agreement in Polarity Identification of Moments by Sentences**

| Average agreement between any 2 evaluators | |
|---|---|
| Average Agreement | Average Weighted Kappa |
| 84.5% | 0.445 |

| Weight Map Used for Weighted Kappa | | | | |
|---|---|---|---|---|
| | | **Evaluator B** | | |
| | | positive | neutral | negative |
| **Evaluator A** | positive | 0 | 1 | 2 |
| | neutral | 1 | 0 | 1 |
| | negative | 2 | 1 | 0 |

Despite the high average agreement we found there to be a relatively low weighted Kappa. This indicates that much of the agreement observed between any two given raters may be due to chance. Part of this could be due to an imbalance seen in the data, with neutral moments constituting the majority of the data analyzed. This may also be a natural observation given the large number of raters used and data to support low affectual agreement between human raters [30]. It is also a possibility that student raters were not as careful with ratings as usability experts may have been. Though there are studies that show non-expert labeling may be as effective as labels created by experts for affect recognition [9], the low kappa score on our dataset and the known disagreement between paper's studying human agreement, indicates the need for more research in regards to the true level of human agreement and to the dataset used. However, for the purposes of this study, an accuracy threshold of 80% was chosen for our analyzers to meet. Analyzers above this threshold in both precision and recall indicate they may be useful tools for the automation of affectual data collection in usability evaluations.

The following section outlines the results of our evaluation and a discussion of the findings. As stated, precision, recall, and F1-scores were calculated for sentences corresponding to moments of positive polarity and negative polarity. However, because text-based analyzers cannot factor in tonal or visual indicators, we also examined their ability to pick up on moments that our human-participants found through visual-only cues, audio-only cues, or both cues. A separate set of precision, recall, and F1-score metrics were calculated for the positive sentiments found through each cue..

Table 6.1 indicates the performance metrics for positive and negative polarity prediction in each of our sentiment analyzers. These statistics present a broad look into the overall ability of each analyzer to detect and correctly classify the same affectual data found by human evaluators.

**Table 6.1 : Performance Metrics for Positive and Negative Polarity Prediction**

| Method | Overall Accuracy | Positive Polarity | | | Negative Polarity | | |
|---|---|---|---|---|---|---|---|
| | | Precision | Recall | F1-score | Precision | Recall | F1-score |
| AFINN | 0.698 | 0.291 | 0.485 | 0.364 | 0.128 | 0.066 | 0.087 |
| SentiStrength | 0.756 | 0.304 | 0.364 | 0.331 | 0.267 | 0.053 | 0.089 |
| Umigon | 0.777 | 0.353 | 0.182 | 0.240 | 0.400 | 0.107 | 0.168 |
| VADER | 0.619 | 0.197 | 0.606 | 0.297 | 0.378 | 0.184 | 0.248 |

Using a baseline of 80% agreement, none of the tested analyzers performed well enough to be considered useful in automating affectual data collection in the context of user testing.

To get a more accurate picture of the results, however, it is important to look at how well each analyzer predicted sentiment based on the method of communication used to display that sentiment. During labeling of the testing dataset, participants marked the type of indicator that clued them into a particular display of sentiment: audio-only cues, visual-only cues, and both cues. The statistics in Table 6.2 show the performance metrics when taking into consideration the method of display used to communicate sentiment.

**Table 6.2 : Performance Metrics for Polarity Prediction by the Method of Display Used to Communicate Sentiment**

| Positive Polarity | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Method | Visual Cues | | | Audio Cues | | | Both Cues | | |
| | Precision | Recall | F1-score | Precision | Recall | F1-score | Precision | Recall | F1-score |
| AFINN | -- | -- | -- | 0.759 | 0.697 | 0.727 | 0.600 | 0.255 | 0.358 |
| SentiStrength | -- | -- | -- | 0.773 | 0.459 | 0.576 | 0.529 | 0.205 | 0.295 |
| Umigon | -- | -- | -- | 0.909 | 0.212 | 0.00 | 0.600 | 0.128 | 0.211 |
| VADER | -- | -- | -- | 0.803 | 0.795 | 0.799 | 0.594 | 0.442 | 0.507 |
| Negative Polarity | | | | | | | | | |
| Method | Visual Cues | | | Audio Cues | | | Both Cues | | |
| | Precision | Recall | F1-score | Precision | Recall | F1-score | Precision | Recall | F1-score |
| AFINN | -- | -- | -- | 0.207 | 0.029 | 0.052 | 0.625 | 0.098 | 0.169 |
| SentiStrength | -- | -- | -- | 0.182 | 0.027 | 0.047 | 0.750 | 0.058 | 0.107 |
| Umigon | -- | -- | -- | 0.000 | 0.000 | 0.000 | 1.000 | 0.170 | 0.290 |
| VADER | -- | -- | -- | 0.229 | 0.057 | 0.091 | 0.800 | 0.231 | 0.358 |

Taking into consideration methods of sentiment display, we found each analyzer to be higher performing for audio cues, undesirable for moments involving both audio and visual cues, and completely inadequate for visual cues. Text-based sentiment analysis techniques have no method of detecting visual or tonal cues and this is reflected in the data. None of the visual-only labeled moments seen by our participants were discovered by any of the tested sentiment analyzers. Though these are intuitive findings, they are also important to note for the context of usability testing. Humans use a variety of methods to express themselves. Sometimes we don’t even realize when we’re getting stressed or frustrated and are unable to verbalize feelings we haven’t yet realized. Other times we may not realize a feeling is important or we may simply forget to verbalize our feelings, especially during a test.

That being said, all of our sentiment analyzers performed better for audio cues. In general each analyzer was more effective at identifying and predicting positive and neutral moments over negative moments. Negations and negative sentiment are historically more difficult for text-based sentiment analysis and our evaluation reflected that.

In regards to positive audio cues, all analyzers showed fairly high precision but only VADER and AFINN had good recall and F1-scores. The recall scores show that moments labeled as positive by human evaluators were also typically labeled as positive by VADER and AFINN. Further the higher precision scores show that when a positive prediction was made, it was usually made correctly. Between VADER and AFINN, AFINN was more likely to predict neutral moments correctly and had higher precision for both neutral moments, but VADER was the only technique with a high enough recall score (0.795) to be able to predict positive moments at the level of a human evaluator.

In regards to negative moments, the predictive abilities of all tested analyzers were surprisingly low. At best, analyzers predicted only 0.5% of negative moments correctly and when a negative prediction was made it was most often wrong. Though negative sentiment is often harder for test-based analyzers to predict, it is likely that the subjects of our mock usability tests and the politeness of our participants further lowered the predictive abilities of each analyzer. In group two, one of the tasks asked participants to “Find what a Blue Alert is on the CHP section of the website”. Sentences with buzzwords like alert, death, and suspect were usually marked as negative because lexicons often list these kinds of words as being highly negative. Lexicons designed for social media also benefit from a larger range of intensity seen in online comments [24]. However, given the more professional, scholastic setting under which our usability evaluations were conducted, most remarks were fairly subdued. Brief comments like “I like the large buttons” or “I don’t know what I should click on …” were more common in our users’ feedback. These contextual differences in usability testing likely contributed to the low performance observed in text-based sentiment analysis of negatively polarized moments.

## Chapter 7 Conclusion

Usability testing, or user experience (UX) testing is an important part of the user interface design process but it can be expensive in terms of time, resources, and consistency. This paper presents preliminary work into one potential solution to these drawbacks: the automation of qualitative data collection through text-based sentiment analysis.

We presented a performance evaluation to test the ability of four well-regarded sentiment analysis methods to correctly detect and classify the polarity of affectual data. Based on our evaluation, we believe that publicly available text-based sentiment analyzers are not in a place yet to provide useful tools for the context of user testing - at least not on their default settings. The only text-based analyzer that met our threshold of 80-85% accuracy was VADER and it was only in the subset of positive moments displayed through audio-only cues. Still, this high accuracy and the decent overall accuracies of each analyzer gives us hope. The social media contexts for which most lexicons have been created do hold significant differences from the usability testing domain and though text-based methods have little ability to recognize visual or tonal cues, many other technologies today do. By designing feature sets specific to user testing domains and incorporating multimodal sentiment analysis tools we may yet accomplish the automation of affectual data collection in usability evaluations.

## Chapter 8 Limitations and Future Works

The application of sentiment analysis to usability testing remains a widely untouched field of research. The findings in this paper only just begin to touch the surface and there exists many more faucets to dive into.This section provides a brief overview of the limitations in our study and lists avenues for extensions and future works.

### 8.1 Speech-to-Text Recognition

Early in the project, there was an attempt to incorporate open-source speech-to-text translators. These would've been used to automatically transcribe user tests into a usable input medium for text-based sentiment analyzers. However, a common issue with automatic transcription, and one we experienced, is that background noise, coarticulation, accents, slang, and homophones often do not translate well [21]. Though a poor transcription could highly limit the usefulness of text-based sentiment analysis, the role of speech-to-text translators are not incorporated into this paper.

### 8.2 Usability Testing Datasets

The mock usability tests used in this study covered a relatively small domain and were not designed by usability experts. Despite our best efforts, a publicly available dataset of usability recordings could not be found. Generating a gold-standard user testing dataset would be highly valuable to future usability research. However, in regards to this project a larger testing set of usability methods and user interfaces would have been ideal.

With regards to COVID-19, only remote usability tests could be conducted and, within that subset, only unmoderated testing methods were employed in this study. To better grasp the extent to which sentiment analysis can be applied to usability testing, a larger dataset could incorporate a wider range of user testing methods, target user interfaces, and studied participants.

### 8.3 Adjusting Sentiment Analysis Parameters

In this study we chose to test all methods of sentiment analysis using their default settings. Though our results were not as promising as we expected, making adjustments to the settings or standard implementations of methods to better accommodate usability contexts could provide much better results. For example, participants often used placeholder words when thinking aloud. Some of these words, like 'okay' and 'well' are listed with positive polarities under the VADER lexicon, despite being predominantly used in a neutral way by participants. Adjusting the related polarity of these placeholder words could result in fewer false positives for the VADER sentiment analyzer [20]. Similarly adjustments could be made to almost all our tested methods to attempt to make tools that are better suited to user testing.

This idea could be taken a step further with the creation of a usability-specific sentiment analysis tool. Most of the analyzers tested were rule-based methodologies that define features and lexicons by which to classify words and sentences [14, 18, 20]. For each tool tested, these features were designed with a specific context and dataset in-mind, cheifly Twitter or Social Media. Though each analyzer was found by Sentibench to be relatively high performing in contexts of social media, movie reviews, and article comments, all of these domains present inherent differences from user testing contexts.

In most usability tests participants engage in one-to-one communication or do not communicate with another person at all. However, on social media users typically engage in short one-to-many communication techniques. This has led to a higher prevalence of summarized speaking, involving acronyms, word reductions, letter/number homophones, stylized spelling, emoticons, and unconventional/stylized punctuation [23]. While these can be important affectual data in the context of social media, they are unlikely to be present or relevant in transcriptions of usability tests. The method of communication, being verbal in user testing and written in social media contexts, can also impact features of our language. Written language tends to have greater lexical diversity, more difficult words, simpler sentences, greater idea density, and a lower verb to adjective ratio [22]. Further, various studies have suggested that prompting communication, private communication, and face-to-face communication, can all have effects on the linguistic structure of our responses [23, 24]. Customizing features to account for these differences could improve the use of text-based sentiment analysis in the context of user testing and in other contexts as well.

### 8.4 Multimodal Sentiment Analysis

Though text-based sentiment analysis presents many unique advantages to detecting affectual data, there are drawbacks. Text-based sentiment analysis lacks the ability to factor in visual and tonal cues that indicate the presence of sentiment. In the context of usability testing, moments when participants forgot to think-aloud but displayed things like furrowed brows or gasps went undetected by our sentiment analyzers. These could be important indicators in differentiating things like a positive, affirmative 'okay' and neutral, placeholder 'okay'. Other forms of sentiment analysis may be able to provide better feedback in these cases, like facial recognition and speech analysis. Implementing multimodal methods of sentiment analysis may provide a better account of the many-faceted ways humans communicate emotions and opinions.

## References

[1] Melody Y. Ivory and Marti A Hearst. 2001. The State of the Art in Automating Usability Evaluation of User Interfaces. ACM Comput. Surv.33, 4 (Dec. 2001), 470–516. https://doi.org/10.1145/503112.503114

[2] C. Karat, "Cost-Benefit Analysis of Usability Engineering Techniques," Sage Journals, [Online]. Available: https://journals.sagepub.com/doi/abs/10.1177/154193129003401203. [Accessed 10 January 2020].

[3] Filipe N Ribeiro, Matheus Araújo, Pollyanna Gonçalves, Marcos André Gonçalves, and Fabrício Benevenuto. 2016. Sentibench-a benchmark comparison of state-of-the-practice sentiment analysis methods. EPJ Data Science5, 1 (2016), 1–29.

[4] G. Chao. 2021. Applying Facial Emotion Recognition to Usability Evaluations to Reduce Analysis Time. Master's Thesis, California Polytechnic State University, San Luis Obispo. DigitalCommons@CalPoly. (2021)

[5] S. M. Mohammad. Sentiment analysis: Detecting valence, emotions, and other affectual states from text. In *Emotion measurement*, pages 201–237.Elsevier, 2016.

[6] A. Tumasjan, T. Sprenger, P. Sandner, and I. Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. InProceedings of the International AAAI Conference on Web and SocialMedia, volume 4, 2010.

[7] J. Nielsen. Thinking aloud: The #1 usability tool, Jan 2012. https://www.nngroup.com/articles/thinking-aloud-the-1-usability-tool/

[8] M. Hertzum, P. Borlund, and K. B. Kristoffersen. What do thinking-aloud participants say? a comparison of moderated and unmoderated usability sessions. International Journal of Human–Computer Interaction, 31(9):557–570, 2015.

[9] Snow R, O'Connor B, Jurafsky D, Ng AY (2008) Cheap and fast - but is it good?: evaluating non-expert annotations for natural language tasks. In: Proceedings of the 2008 conference on empirical methods in natural language processing (EMNLP '08)

[10] K. Moran. Remote usability-testing costs: Moderated vs. unmoderated, Jul2020. https://www.nngroup.com/articles/remote-usability-testing-costs/

[11] Shahnawaz and P. Astya. Sentiment analysis: Approaches and open issues. In2017 International Conference on Computing, Communication andAutomation (ICCCA), pages 154–158, 2017.

[12] Nielsen F (2011) A new ANEW: evaluation of a word list for sentiment analysis in microblogs. arXiv:1103.2903

[13] Bradley MM, Lang PJ (1999) Affective norms for English words (ANEW): stimuli, instruction manual, and affective ratings. Technical report, Center for Research in Psychophysiology, University of Florida, Gainesville, FL

[14] F.˚Arup Nielsen. afinn, 2019. https://github.com/fnielsen/afinn

[15] Thelwall, M., Buckley, K., & Paltoglou, G. (2012). Sentiment strength detection for the social Web, Journal of the American Society for Information Science and Technology, 63(1), 163-173.

[16] M. Thelwall. Sentistrength access page, 2021. http://sentistrength.wlv.ac.uk/

[17] C. Levallois. Umigon: sentiment analysis for tweets based on terms lists and heuristics. InSecond Joint Conference on Lexical and ComputationalSemantics (*SEM), Volume 2: Proceedings of the Seventh InternationalWorkshop on Semantic Evaluation (SemEval 2013), pages 414–417,Atlanta, Georgia, USA, June 2013. Association for ComputationalLinguistics.

[18] C. Levallois. Umigon nocode app, 2021. https://seinecle.github.io/nocodeapp-mods/

[19] C. Hutto and E. Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. InProceedings of the InternationalAAAI Conference on Web and Social Media, volume 8, 2014.

[20] C. Hutto. Vadersentiment, 2020. https://github.com/cjhutto/vaderSentiment

[21] V. Prasad. Voice recognition system: Speech-to-text.Journal of Applied andFundamental Sciences, 1(2):191, 2015.

[22] W. Chafe and D. Tannen. The relation between written and spoken language.Annual review of anthropology, 16(1):383–407, 1987.

[23] D. Barton and C. Lee. Language online: Investigating digital texts and practices. Routledge, 2013.

[24] V. Kulkarni, M. L. Kern, D. Stillwell, M. Kosinski, S. Matz, L. Ungar,S. Skiena, and H. A. Schwartz. Latent human traits in the language of social media: An open-vocabulary approach.PloS one, 13(11):e0201703,2018.

[25] P. Barba. Sentiment accuracy: Explaining the baseline and how to test it, Aug2020. https://www.lexalytics.com/lexablog/sentiment-accuracy-baseline-testing#:~:text=When%20evaluating%20the%20sentiment%20(positive,training%20a%20sentiment%20scoring%20system.

[26] T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. InProceedings of human language technology conference and conference on empirical methods in natural language processing, pages 347–354, 2005.

[27] P. Ranganathan, C. Pramesh, and R. Aggarwal. Common pitfalls in statistical analysis: Measures of agreement.Perspectives in clinical research, 8(4):187, 2017.

[28] J. R. Landis and G. Koch. The measurement of observer agreement for categorical data. Biometrics, 33 1:159 – 74, 1977.

[29] Zaiontz, C. (2014). Chris Heard. Real Statistics Using Excel. https://www.real-statistics.com/reliability/interrater-reliability/weighted-cohens-kappa/.

[30] R. Molich, M. R. Ede, K. Kaasgaard, and B. Karyukin. Comparative usability evaluation. Behaviour & Information Technology, 23(1):65–74, 2004.