Sentiment analysis tool comparison using Vader and SentiStrength

Miika Pernu, Mikko Roimaa, Pilvi Tunturi

University of Oulu, firstname.lastname@student.oulu.fi

*Abstract* - This project aims to investigate the sentiment polarity (positive, negative, neutral) and intensity of hotel reviews and compare it to human coded ratings. Correlation between the human ratings and sentiment intensity by the tool can be regarded as evidence that the analysis is correct. The interesting part is not that much sentiment itself, but rather argumentation to the polarity. We will detect Pearson’s correlation coefficient values by comparing SentiStrength and NLTK Vader sentiment analysis tools and aim to find contrast evidence from individual raters. We will compare how well these tools are able to correlate the sentiments of user’s reviews and study what are the strengths and weaknesses of those tools and how do they perform overall compared to each other and together. We will analyze what kind of categories relate to each review and how do they correlate to positive or negative sentiments. We will also investigate what kind of named entities each review has and how do their presence correlate to the positive or negative sentiments.

During this study, several hypothesis will be tested. Firstly hypothesis on the relation of argumentation on positive and negative sentiments. Secondly we will test hypothesis related to review ambiguity. Are ambiguous reviews shorter or do they have bad readability.

*Index Terms* – Comparison, opinion mining, sentiment analysis, tool.

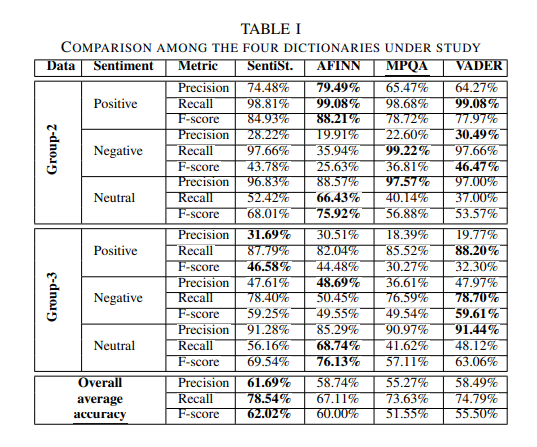
Introduction

With web, social media and digitalization the growth in user generated content has opened a huge opportunity for businesses to find new ways to generate value from data. Since the early 2000s, opinion mining or sentiment analysis has become one of the most researched area in natural language processing. [1, 2] As sentiment can be seen as an attitude or a thought as a response to a feeling [1], it is interesting in context like customer satisfaction and NPS surveys. However, the more interesting are the facts related to the sentiments [10]. The nature of sentiment analysis is interdisciplinary. It can be approached from computer science point of view, social science and for example management science [2].

This study investigates opinion mining or sentiment analysis mainly from the lexicon based tooling perspective. Similar studies have been made with these and other lexicon based tools with similar and different datasets. [6,7] However, neither of these studies look into argumentation as such, which is important for any business to draw conclusions on their products and services.

For businesses nowadays it is not difficult to start analyzing sentiments of their customers in their social media channels, customer support and service channels or customer reviews and feedbacks. Especially for English. The challenges with tooling can be many though. When choosing the tool, it is important to understand the purpose. [3, 4] Using sentiment analysis for social media posts requires different tool than running a tool used for news analysis. [4]

One must also understand other parts, such as developer community, programming language, portability and other aspects typically evaluated when buying similar tools. One must also consider the usage patterns of the tooling. Is it going to be used off the shelf with standard configuration, or is it going to be trained with data. [3] There are similar studies, that have compared Vader and SentiStrenght approach [6,7]. Study by Al-Shabi [7] analysed 5 most important lexicon based sentiment analysis tools, including Vader and SentiStrength. The study revealed that classification with Vader was more accurate among negative and positive sentiments. The study was conducted with Twitter data. We will conduct this study using two tools, Vader and SentiStrenght with data from Kaggle containing ten thousand hotel reviews in English. The data will be used as is without largely preprocessing it. Study by Zibran [6] analyzed 4 lexicon based tools, including SentiStrenght and Vader. The study used domain specific dictionaries to improve on the accuracy. The performance of the tools in dedicated or special domains such as Software Engineering was known to be less accurate. The study concluded that “We have studied the prospect and effectiveness of four distinct dictionary building methods for sentiment detection in SE texts. Based on a quantitative analysis over a benchmark dataset, we have found that dictionaries (i.e., SentiStrength, AFINN, and VADER) created using simple lexicon-based approach perform better (for SA in SE text) than those (i.e., MPQA), which include complex techniques for incorporating subjectivity and contextual sense.” However, this study did not conclude a clear winner, but SentiStrength was seen performing better with negative sentiments.



[HUOM MUISTA VIITATA TUONNE OIKEIN]

An empirical analysis by Singha et al [5] showed that there is a high correlation between customer ratings and sentiments. These types of findings set a stage for our study too. MAINITAAN MEIDÄN KORRELAATIO RESULTSEISSA. SentiStrength: Vader:

The study by Thelwall et all also shows that SentiStrenght is performant enough with different types of social web texts [9]. But the argumentation analysis is needed as a step forward from sentiment analysis to determine the impact of a certain review [10].

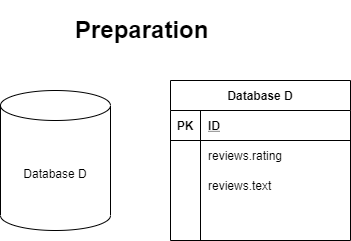
METHODOLOGY

Our study concentrated on English hotel reviews data from Kaggle.com. The analysis flow can be illustrated with following diagram:

Diagram

Description automatically generated

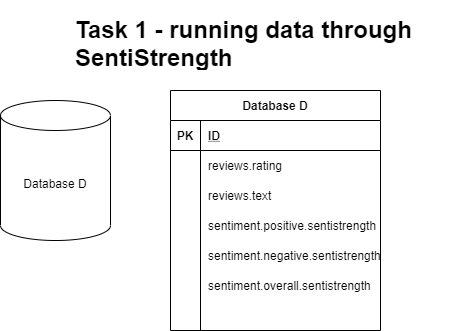
We started by downloading the data and manually inspecting it. We became aware that dataset contained hotel responses within the reviews, but we could not find a way to clean and process those ten thousand reviews. Hotel reviews in nature are such that they can be longer or shorter. We acknowledged that the impact of those responses within data to be analyzed makes the analysis less trustworthy and too positive, but we agreed that within that amount of source data, the impact is not significant [ollaanko tätä mieltä? ]. The study commences that the data collection process has been erroneous but there is enough data to determine results. We investigated 3 different datasets and chose the one with more garbage reviews. With cleaner ones, results seemed too clean, so we decided to choose the one with more variance in the data. It is easier to analyze readability and whether a review is considered badly written. A review is badly written, if the percentage of recognized words is low. Finally, the data was cleaned up by removing unnecessary columns to speed up the process. Reviews.text and reviews.ratings were taken as is and also review.id was constructed. With this step, encoding was ensured to be utf-8 throughout the project.



The next step in the process was to run data through SentiStrength.

The SentiStrength dictionary is constructed by combining LIWC and GI dictionaries similar to VADER, and also includes lists of emoticons, negations and intensifiers. SentiStrength is a sentiment analysis tool that performs with human level accuracy in English social media texts [8]. It is lexicon based, designed to give a strength to a term. For example “love” has a stronger positivity than “like” [9].

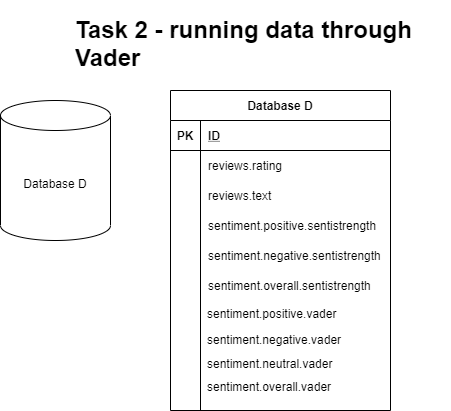
Data was fed into SentiStrength java client. That was seen more performant than the windows client. In this task each review was run through SentiStrength. Data was first processed into a format which is recognized by SentiStrength (tab separated CSV) after which SentiStrength assessed the sentiment polarity for each review individually and wrote it to the file parsed for SentiStrength. After analysis was completed the input file for SentiStrength was read and combined with the main database, overall sentiment was calculated as additional step and temporary file was deleted as it was now obsolete.



We also ran the data through AWS Comprehend, which is a tool offered by Amazon to find insights and relationships from data using machine learning techniques. Setting up AWS comprehend to analyze the dataset was easy, but due to the cost it was decided not to be used in this analysis. While the cost seemed bearable, the unpredictability of the billing rules was seen as a threat to the project.

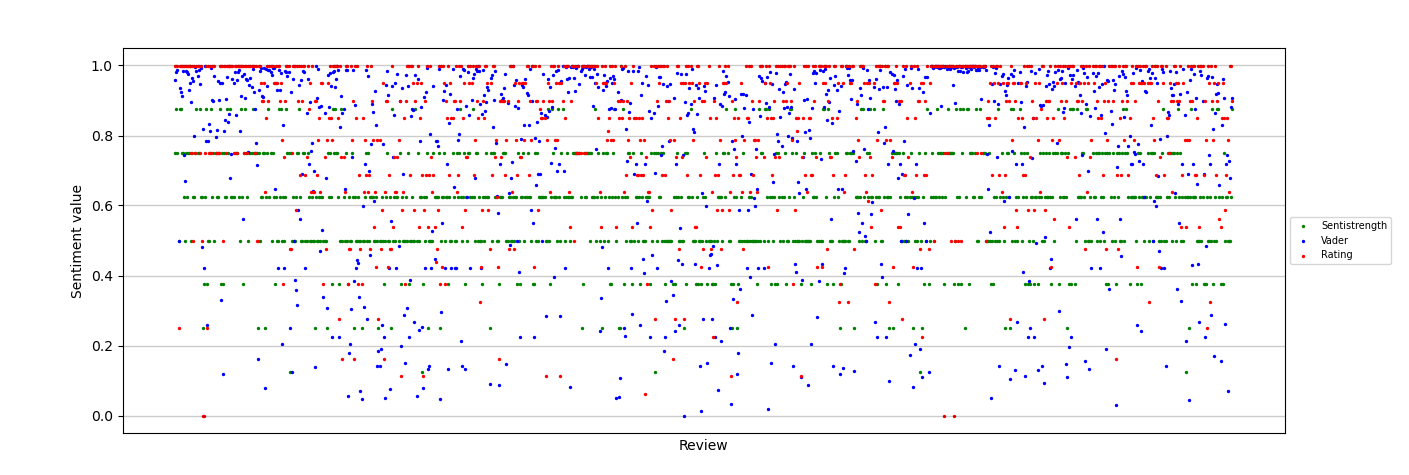
To replace AWS Comprehend, we decided to use NLTK Vader. Vader by Gilbert et al [11] in their study present and evaluate Vader, Valence Aware Dictionary and sEntiment Reasoner. Vader is a simple lexicon and rule-based model for sentiment analysis. It is specifically attuned to polarity and intensity of sentiment expressed in social media texts. It works well on texts from other domains too [12]. The study by Gilbert et al [11] revealed that based on correlation coefficient, Vader performs and even outperforms individual human raters in classifying polarity. It is proven that Vader sentiment lexicon is gold-standard quality and has been validated by humans.

This task was almost identical with previous one, except that the results were written into the database file directly.

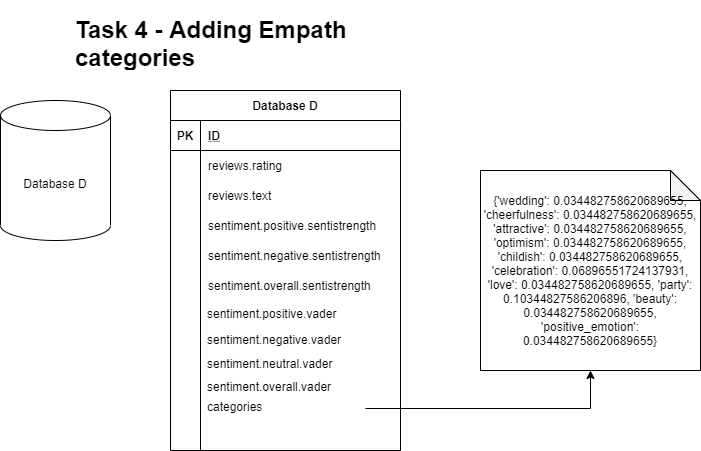


With this data, the Vader and SentiStrength results were plotted in the same graph along with the actual reviews. The results from both analyzers were normalized along with the actual reviews to get a more meaningful graph. The graph contains only the first 1000 points of the dataset as plotting the results for all 10 000 values wouldn't look very informative at all.

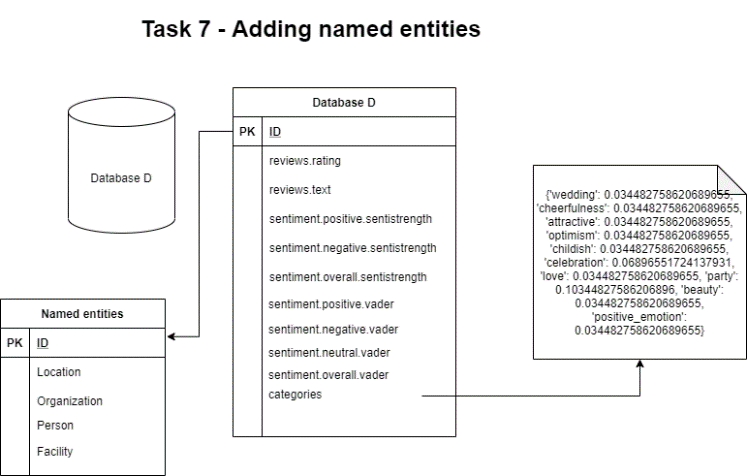
Additionally pearson coefficient correlation is calculated for both analyzer results in relation to the user ratings.



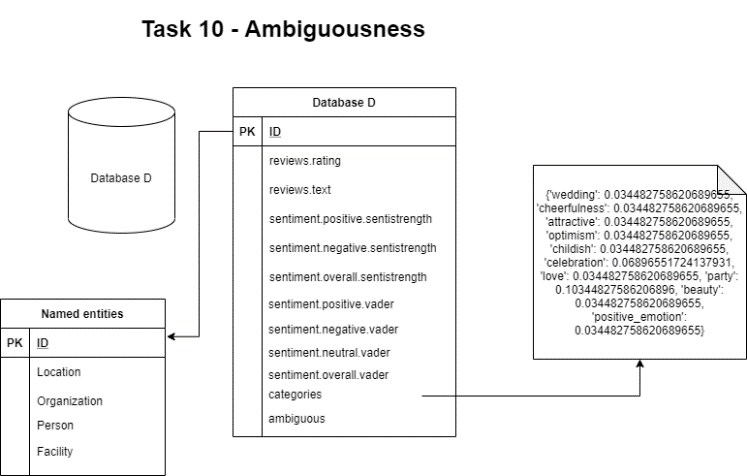
For the preprocessed, parsed and analyzed data, we started to enrich it with lexical categories. We decided to use Empath tool for it. Empath is a lexicon mined from modern texts. It groups words into topics, and is human validated. Empath uses the combination of machine learning and crowd sourcing. Compared to the well established LIWC categories, Empath is much wider. It contains over 200 categories compared to the 40 categories of LIWC. And as our language evolves, so does Empath. [14] Empath categories were constructed by Python library and written into the reviews database. Additionally each unique category was extracted and saved into empath\_categories.txt separately.

Categories were stored as key-value pairs into the Database D. 

Next step was to include named entities into the database. The problem we observed with nltk named entities was that it’s recognition capabilities are limited. It recognized capitalized nouns as “persons”. The places, like Best Western, were recognized as two different entities, one as a name (Best), other as a organization (Western). We decided to use binary format on whether the entity is in place or not for the review. The given type presence was associated with sentiment polarity to analyze the potential correlation between those. The ratio of certain types of named entities vs. negativity & positivity was analyzed.



Task 10 and 11 were about the ambiguousness of the review. Each review was split into one of two classes: ambiguous or non-ambiguous. Whether a review belonged in the ambiguous class was determined by whether sentiment analyzer VADER result has significant deviation from the users own rating. Additionally in task 10 it was tested whether reviews in ambiguous class were likely to be badly written. Whether a review is badly written is determined by the percentage of known words in the review. A word is considered known if WordNet is able to find any synsets for the word. A word without synsets is considered unknown.



Finally for task 11 it was checked whether ambiguous reviews are likely to be shorter than others. The result was printed into the application GUI.

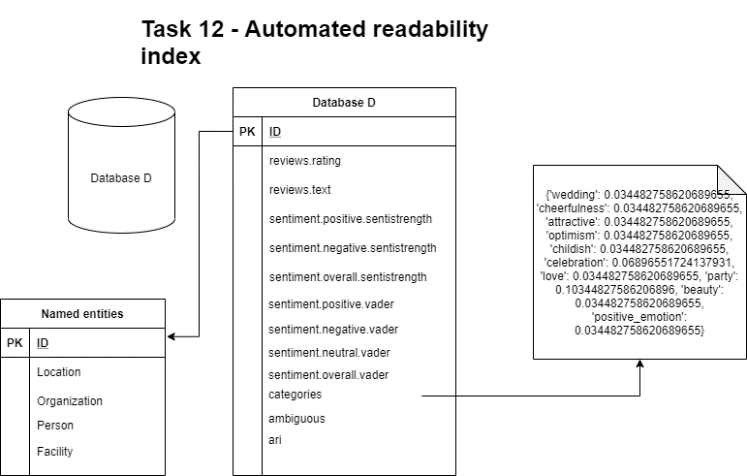
Average known words percentage for ambiguous: 0.6558287948650917

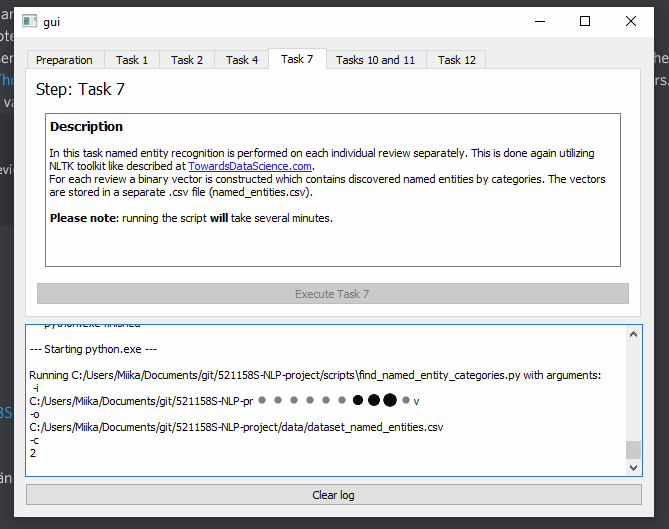
Average known words percentage for unambiguous: 0.6561596003831222

After that the next goal was to test the following hypothesis: ambiguous reviews have bad readability. This hypothesis was tested by calculating the Automated Readability Index (ARI).

ARI is an index designed to measure understandability of English text. It represents approximately a grade level that is needed to comprehend certan text. [15]

We ran ARI calculation for each review to see which class (ambiguous vs non-ambiguous) has the larger value by average. The results were printed into the application output panel. Additionally the ARI value was written into the database.



GUI is build using QT and a user can run the application with simple step by step user interface: 

Figures, Tables, and Equations

All figures and tables must fit either one or two-column width, 3.4” or 7” wide respectively. It is suggested that you use one-column figures and tables whenever possible. If your table or figure will not fit into one column, then insert a continuous section break before and after the table or figure, as described above and define it as one-column. To make the paper read easier, you may want to position any table or figure that requires the full width of the paper either at the bottom of the page or the top of a new page.

Do not abbreviate “Table” or “Figure.” Use Roman numerals FOR BOTH. Use the following format guidelines for figures and tables:

* **Figure and table headings**: 10 point, Times New Roman UPPERCASE, centered. Place below the figure and above the Table (this style is defined under the style menu of this document as “Figure Heading”). Use Roman numerals.
* Leave one blank line above and below each Table or Figure.
* **Figure and table captions**: 8 point, Times New Roman, Small Caps, centered. Place below the figure or table headings (this style is defined under the style For task 10 each review is split into one of two classes: ambiguous or non-ambiguous. Whether a review belongs in the ambiguous class is determined by whether sentiment analyzer VADER result has significant deviation from the users rating.  
    
  Additionally in task 10 it is tested whether reviews in ambiguous class are likely to be badly written. Whether a review is badly written is determined by the percetange of known words in the review. A word is considered known if WordNet is able to find any synsets for the word. A word without synsets is considered unknown.  
    
  Finally for task 11 it is checked whether ambiguous reviews are likely to be shorter than others.sure you use title case.
* **Table text**: 8 point, Times New Roman, (this style is defined under the style menu of this document as “Table text”)

Table I and Figure I below illustrates proper Table and Figure formatting. Avoid placing figures and tables before their first mention in the text. IEEE has the following rules for inserting graphics as figures:

* The manuscript’s graphics should have resolutions of 600 dpi for monochrome, 300 dpi for grayscale, and 300 dpi for color.
* Graphics should be inserted into the manuscript file by clicking on “Insert – Photo – Picture From File.” This means you must save every graphics as a separate file. Do not use cut and paste to insert graphics.
* Do not link to a graphic. When inserting figures or tables be sure you insert the figure and not just a link to the figure. The best way to make sure you are doing this correctly is to save your paper, then open the file on a different machine and make sure all your figures are correct. If you insert the link instead of the figure or table, a box with a big red x will appear in the location where the table or figure is supposed to be located.
* **DO NOT use text boxes for forcing in a table or figure that needs the full width of the paper.**
* **DO NOT use text boxes for captions.**

TABLE I

Point Sizes and Type Styles

|  |  |  |
| --- | --- | --- |
| Points | Type of Text | Type Styles |
| 8  10  8  8  8  10  10  10  10  10  10  11  24 | Table text  Figure and Table Headings  Figure and Table Captions  Footnote  Reference list  Abstract  Index Terms  Section Titles  Main Text and Equations  Subheadings  Author email  Author name  Title | UPPERCASE  Small Caps  **Bold**  **Small Caps, Bold**  *Italic*, Left justified  Title Case |



Figure I

Logo of the Institute for Electrical and Electronics Engineers

Number equations consecutively with equation numbers in parentheses flush with the right margin, as in (1). This is best achieved by using a right tab. DO NOT USE SPACES to position your equations.

*2jk ∂u/∂z = ∂2u/∂x2 + k2 (n2* - β*2) u*  (1)

Refer to “(1),” not “Eq. (1)” or “Equation (1),” except at the beginning of a sentence: “Equation (1) is….”

Make sure you use only the “Symbol Font” for all your symbols, or embed all your different symbol fonts within the file when you save the document.

Headers and Footers

Authors do not add header and footer information. The Publications Chair will add the standard IEEE headers and footers as part of preparing the papers for publication.

Paper Length Limit

ISEC has a hard limit of **8 pages per full paper, and 4 pages per WIP paper**. Papers cannot exceed this length. **Authors CANNOT alter margins, paragraph spacing, or font sizes to make longer papers fit this limit.** Papers longer than 8 pages will be returned to the authors for rewriting.

Template Use

**DO NOT ALTER THE TEMPLATE.** Authors who change margins and font sizes, do not use the requested figure and table title formats, do not use the requested reference format, or otherwise do not use the template will have their paper returned to them for correction.

Formatting reminders: the first paragraph in a section or subsection is not indented; subsequent paragraphs are first line indented at .25”. Section headers have spacing of single with 8 pt before and after. Subsection headers have spacing of single with 6 pt before and after. Don’t have spaces between paragraphs.

Other formatting issues that will result in your paper being sent back to you to re-edit include having tables breaking over columns, (sub) headings being separated from the related text over column or page breaks, the use of hard returns (the enter key) to force a column or page break (use Insert, Break, then either Page or Column).

Acknowledgment

The preferred spelling of the word “acknowledgment” in American English is without an “e” after the “g.” Use the singular heading even if you have many acknowledgments. Please put the sponsor acknowledgments in this section. Do not use a footnote on the first page.

IEEE Standards: No Live Links and Use Embedded Fonts

Do not have live links (URLs) in your paper. Remove the link (right click, edit hyperlink, remove link) and then the http:\\or https:\\ term. All fonts must be embedded. Embedding typically occurs during the creation of a PDF. IEEE requires PDF version 1.7.

Copyright Form

ISEC uses the IEEE electronic copyright form. EDAS will provide a link to the form when you upload your final paper. You MUST sign the copyright over to IEEE to have your paper included in the proceedings.

You MUST receive any client approvals of your paper as soon as possible to avoid proprietary/IP issues that could prevent your paper from being presented at ISEC.

Present to Publish

You MUST present your paper at ISEC for it to be included in the proceedings submitted for archival in IEEE Xplore. You cannot have a non-author present your paper.

In Text References

All material from any research resource must be accompanied by a bracketed in-text reference. This reference must correspond to its end-text full bibliographic information in the References section. Failure to properly reference all resource material used in a paper leaves the paper’s author open to charges of plagiarism.

Follow these specifications for **in-text references**:

* Bracket all in-text references: for example, [1].
* In text references must be **numbered sequentially in the text, beginning with [1]** for the first reference. In other words, the first source from which you quote, paraphrase or use information must be referenced in your paper as [1]. The next source from which your quote, paraphrase or use information must be [2]. If in later in your paper, you use information from the same source and same page as [1], then your in-text reference number will again be [1].
* Do not say “Ref. [3]” or “reference [3].” Simply use the bracketed number thusly: [3].
* For material summarized from several sources, use the appropriate bracketed numbers, for example [3]-[5].
* Bracketed reference numbers should appear after the quotation marks on an in-text quote, but before the final punctuation of the quote. For example, “Here’s the quote” [3]. Bracketed references for paraphrases or summaries should appear after the paraphrase or summary, but before the final punctuation of the sentence or passage. For example, Here’s the paraphrased material [4].

References

Place references in a separate References section at the end of the paper. Number the references sequentially by order of appearance, not alphabetically. List up to three authors’ names in a reference; replace the others by “*et al*.”

[1] Fan, Xing and Zhan, Justin June 2015 “Sentiment analysis using product review data.” 2015.

https://journalofbigdata.springeropen.com/articles/10.1186/s40537-015-0015-2 Web. Accessed: November 1, 2020.

[2] Liu, Bing 2015 “Sentment Analysis: Mining Opinions, Sentiments, and Emotions”. Second edition, preface xi.

[3] Alexandre Pinto, Hugo Gonçalo Oliveira and Ana Oliveira Alves 2016 “Comparing the Performance of Different NLPToolkits in Formal and Social Media Text ”https://drops.dagstuhl.de/opus/volltexte/2016/6008/pdf/OASIcs-SLATE-2016-3.pdf

[4] Aue, Anthony and Gamon, Michael January 2015 “Customizing Sentiment Classifiers to New Domains: A Case Study”.

[5] M.Geetha, Pratap, Singha, SumedhaSinha “Relationship between customer sentiment and online customer ratings for hotels – an empirical analysis. “

[6] Islam, Rakibul and F. Zibran, Minhaz 2017 “A Comparison of Dictionary Building Methods for Sentiment Analysis in Software Engineering Text”

[7] Al-Shabi, M. A. January 2020 “Evaluating the performance of the most important Lexicons used to Sentiment analysis and opinions Mining”

[8] <http://sentistrength.wlv.ac.uk/>, Accessed: November 1, 2020.

[9] Thelwall, Mike, Buckley, Kevan, Paltoglou, Georgios 2012, “Sentiment Strength Detection for the Social Web” <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.278.5294&rep=rep1&type=pdf>

[10] Wachsmuth, Henning, Trenkmann, Martin, Stein, Benno, Engels, Gregor, Palarkarska, Tsvetomira 2014 “A Review Corpus for Argumentation Analysis”

[11] Gilbert, Eric and Hutto C.J 2014 "VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text"

[12] https://pypi.org/project/vaderSentiment Web. Accessed November 1, 2020

[13] https://www.ijcai.org/Proceedings/2017/0677.pdf

[14] Bernstein, Michael S., Binbin, Chen, Fast, Ethan 2017 "Lexicons on Demand: Neural Word Embeddings for Large-Scale Text Analysis" https://www.ijcai.org/Proceedings/2017/0677.pdf

[15] Automated readability Index, web. Accessed November 5th 2020. https://en.wikipedia.org/wiki/Automated\_readability\_index

Author Information

**Mikko Roimaa,** student. Faculty of Information Technology and Electrical Engineering, University of Oulu.

**Miika Pernu**, student. Faculty of Information Technology and Electrical Engineering, University of Oulu.

**Pilvi Tunturi**, student. Faculty of Information Technology and Electrical Engineering, University of Oulu.