Sentiment analysis tool comparison using Vader and SentiStrength

Miika Pernu   
University of Oulu  
miika.pernu@student.oulu.fi

Mikko Roimaa  
University of Oulu  
mikko.roimaa@student.oulu.fi

Pilvi Tunturi  
University of Oulu  
pilvi.roimaa@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, ambiguity of reviews and what entails to that and e.g. the impact of named entities. This project 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. This project will also analyze what kind of categories relate to each review and how do they correlate to positive or negative sentiments. It 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 hypotheses will be tested: Presence of positive/negative sentiment associated Empath category in the review entails a positive/negative sentiment, presence of a given type of named entity entails positive sentiment/negative sentiment, negative sentiment entails more argumentation, badly written reviews are likely to be included in ambiguous class, ambiguous reviews are shorter, ambiguous reviews have bad readability. All of the steps and the comparison of how well these tools can correlate the sentiments of user’s reviews is done by programming an application that can be ran step by step by using simple GUI.

Keywords—SentiStrength, Vader, sentiment analysis, opinion mining

# 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 be 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 point of views [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 investigate argumentation as such, which is important for any business to draw conclusions related to their products, services and customer feedback.

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. It is also important to understand and 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] analyzed 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. This study will be conducted 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 lexicon-based tools outperformed tools that incorporate complex techniques e.g. with subjectivity or contextuality. However, this study did not conclude a clear winner, but SentiStrength was seen performing better with negative sentiments.

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. The study by Thelwall et al 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].

As it is clear, sentiment analysis can be ran by using different kind of tools. As a research problem, this is not unique or novel at all. This study aims to learn the pros and cons of the tools used by concentrating on two tools mainly, SentiStrength and Vader. Secondly, the aim of the study is to educate and use and expand the learned skills in practice.

# Methodology

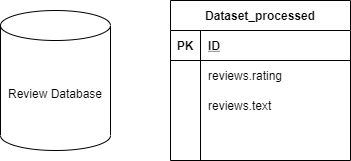
This project study concentrated on English hotel reviews data from Kaggle.com. The analysis flow can be illustrated with following diagram:

Diagram

Description automatically generated

1. Analysis flow

The project started by downloading the data and manually inspecting it. It became clear that dataset contained hotel responses within the reviews, but a way to clean and process those ten thousand reviews could not be found. Hotel reviews in nature are such that they can be longer or shorter and there was not a common pattern to be found for the cleaning process. It was acknowledged that the impact of those responses within data to be analyzed makes the analysis less trustworthy and too positive, but it was agreed that within that amount of source data, the impact is not significant. The study commences that the data collection process has been erroneous, but there is enough data to determine results. Three different datasets were investigated and the one with more garbage reviews was chosen. With cleaner ones, results seemed too clean, so it was 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. 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.

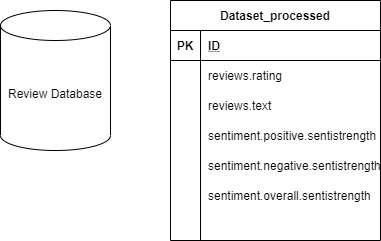


1. Data preparation

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.

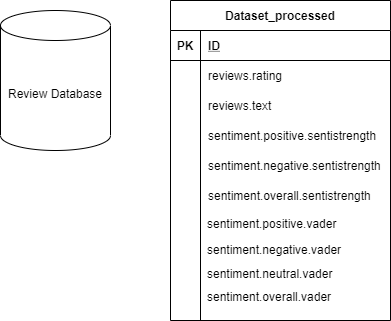


1. Running data through SentiStrength

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 a possible 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.

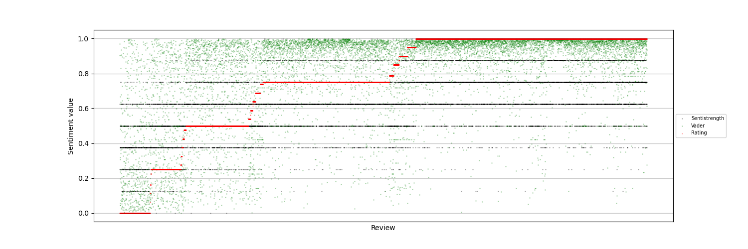
This task was almost identical with previous one, except that the results were written into the database file directly.



1. Running data through Vader

With this data, the Vader and SentiStrength results were plotted in the same graph along with the actual reviews. The data was arranged by the rating. As the ratings were all in different scale, the results from both analyzers were normalized along with the actual reviews to get a more meaningful graph. The graph contains all 10000 points of the dataset.

Additionally Pearson coefficient correlation was calculated for both analyzer results in relation to the user ratings.

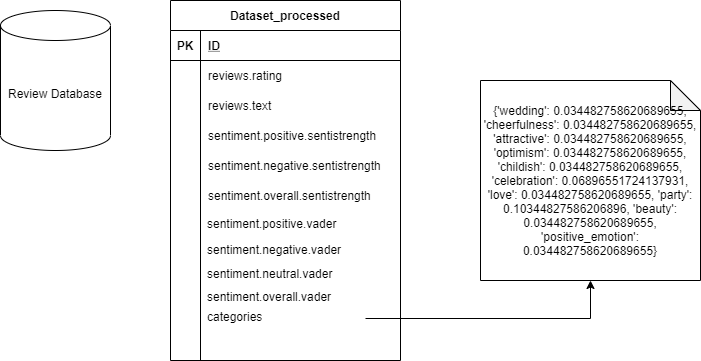


1. Scatter plot showing correlations between user rating and sentiment analysis tools

For the preprocessed, parsed and analyzed data, the process of enriching it with lexical categories was started. Empath tool was the decided 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]

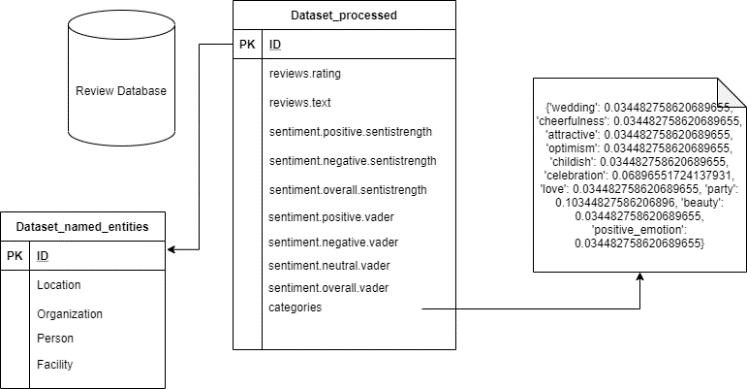
Empath categories were constructed by Python library and written into the reviews database. Additionally, each unique category was extracted and saved into empath\_categories.txt separately.

Categories were stored as key-value pairs into the Review database.



1. Adding Empath categories

Next step was to include named entities into the database. The problem 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 an organization (Western). Binary format was chosen to show 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.



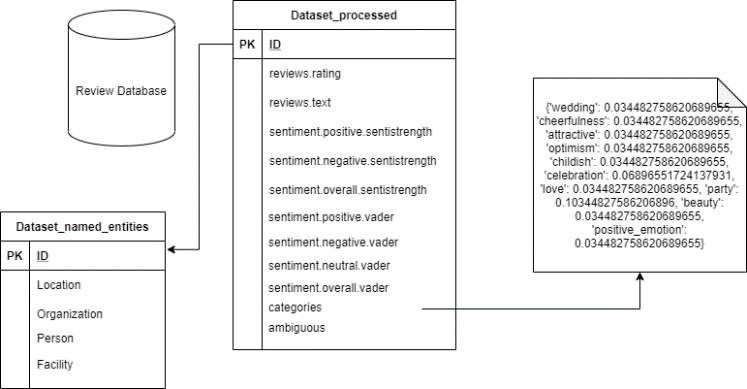
1. Adding named entities

Next step of the project was to test the hypothesis that negative reviews entail argumentation. To be able to accomplish that, a short list of explanation inducing expressions was collected and put together in explanatory\_wording.csv. Each review in the dataset was tested for the number of these expressions found. It is notable that reviews were in the scale of 1-5, positive are >=4 and negative are <=2. Reviews with rating 3 were considered neutral.

Next steps were about the ambiguity of the review.

Resolving ambiguity is one of the biggest problems in NLP. One can see ambiguity in a sentence, if it can be interpreted in two or more ways. [16]

In this process, 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 analyser 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.



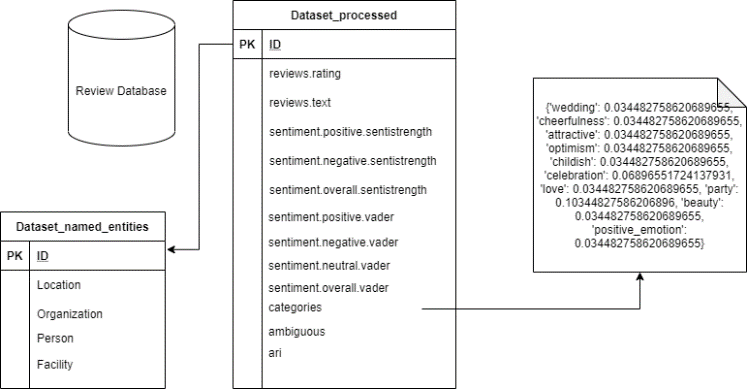
1. Adding ambiguousness value

Finally, it was checked whether ambiguous reviews were likely to be shorter than others. The result was printed into the application GUI.

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).

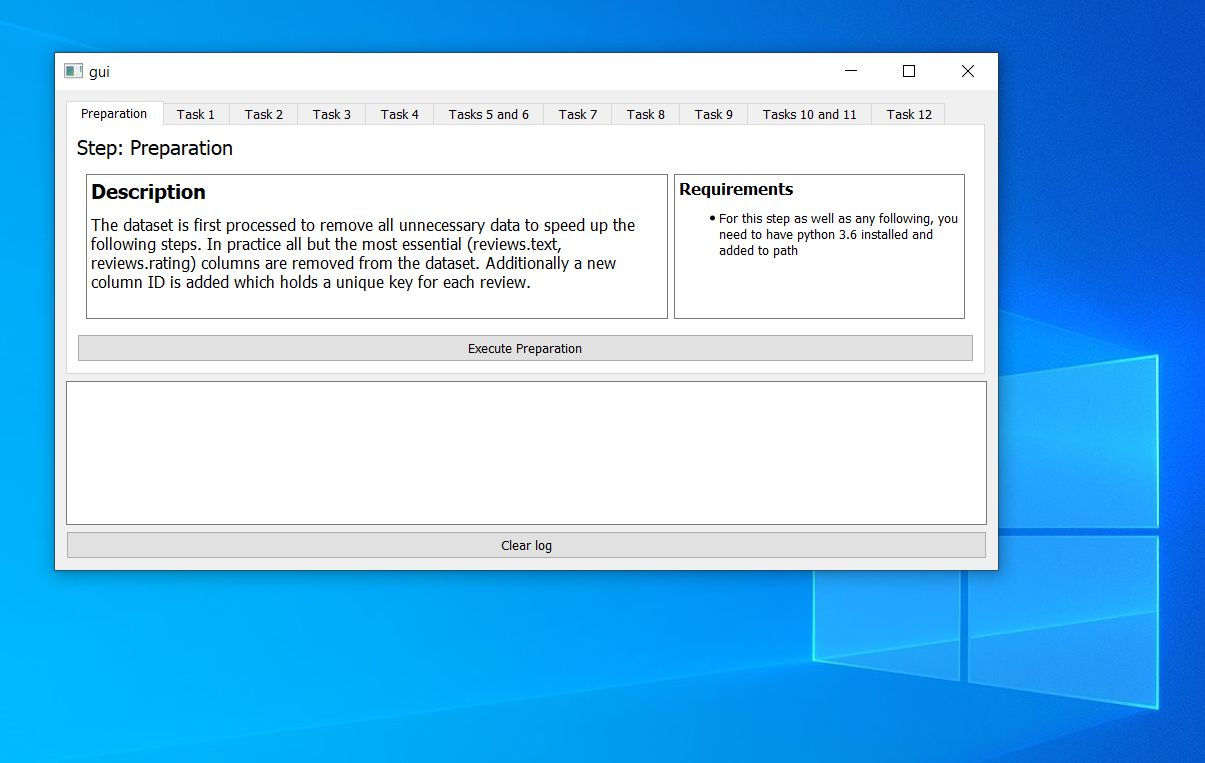
ARI is an index designed to measure understandability of English text. It represents approximately a grade level that is needed to comprehend certain text. [15]

ARI calculation was ran 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.



1. Adding automated readability index (ARI) value

All of the above steps were packaged into the simple graphical user interface (GUI), that was built using QT. User can run the application with simple step by step user interface, that also guides user through the process. The results were saved into review database and results were printed into the output console.



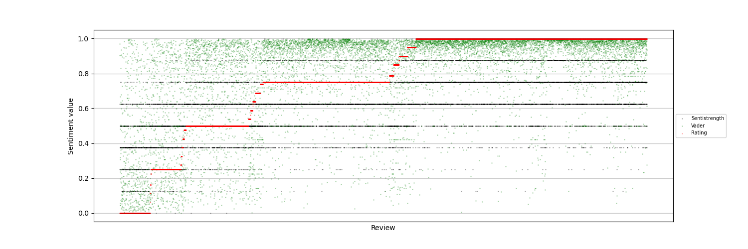
1. Application GUI

# Results

In this chapter, the hypotheses will be walked through, tested and compared with the evidence found from the literature review.

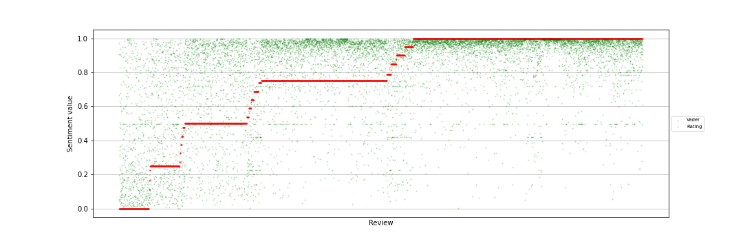
## Plotting the data from sentiment analysis tools

Finding a correlation from big chunks of data can be done using scatterplots. We plotted data from both analyzers into the same graph along with the actual reviews. The data was arranged by the rating. As the ratings were all in different scale, the results from both analyzers were normalized along with the actual reviews to get a more meaningful graph. The graph contains all 10000 points of the dataset.



1. User, SentiStrength and Vader correlation graph

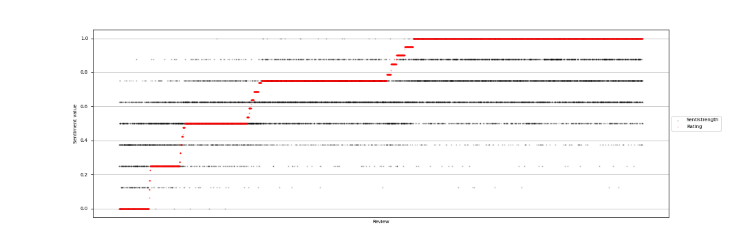
The ten thousand reviews dataset is hard to illustrate in one diagram. Thus, the data was separated per analyzer. The same order was retained. Figure below shows that with Vader, positivity and negativity are best correlated. The correlation in more neutral sentiments is more scattered and not easy to recognize. With SentiStrength, the correlation in negative sentiments is more accurate. This supports also the conclusions made by other studies [6,7]. Pearson correlation coefficient was used to calculate the correlation.



1. User and Vader rating correlation graph
2. Vader pearson correlation coefficient

| VADER Pearson Correlation Coefficient | |
| --- | --- |
| Actual Rating | VADER Analysis |
| 1.000000 | 0.571814 |

1. VADER Pearson Correlation Coefficient

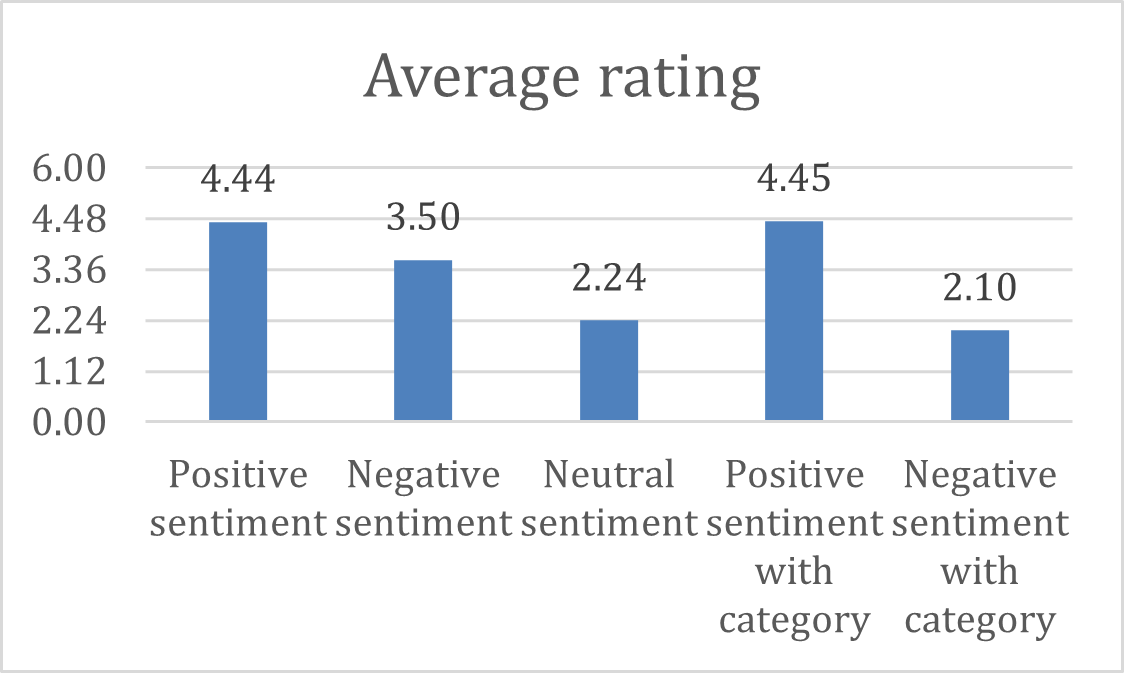


1. User and SentiStrength rating correlation graph
2. SentiStrength pearson correlation coefficient

| SentiStrength Pearson Correlation Coefficient | |
| --- | --- |
| Actual Rating | SentiStrength Analysis |
| 1.000000 | 0.556305 |

1. SentiStrength Pearson Correlation Coefficient

## Hypothesis 1a: “The presence of positive sentiment associated Empath category in the review entails a positive sentiment”



1. SentiStrength Pearson Correlation Coefficient

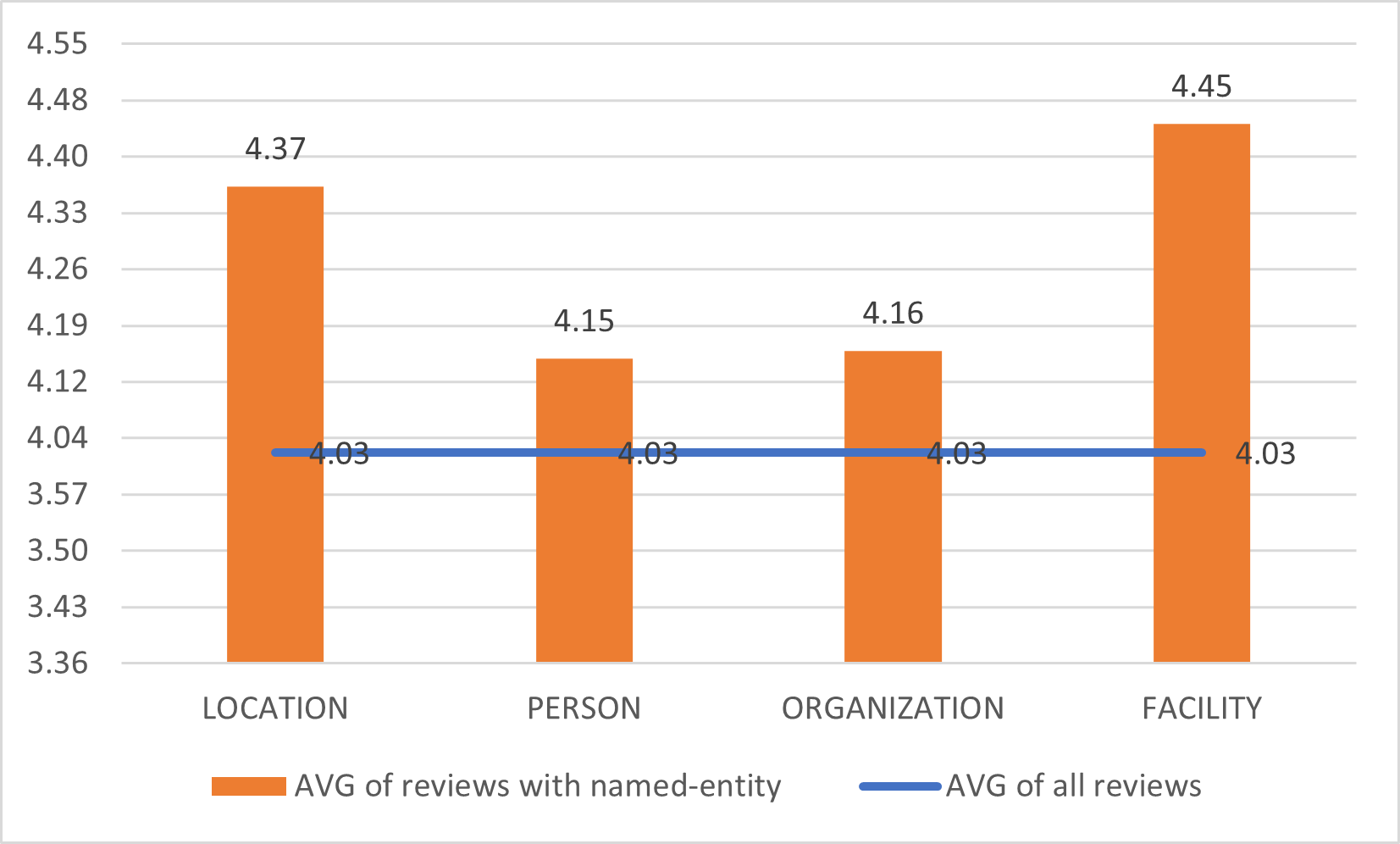
## Hypothesis 1b: “The presence of negative sentiment associated Empath category in the review entails a negative sentiment”.

1. AVG rating

Propose how you will validate such statements using a simple correlation analysis.

## Hypothesis 2: “Presence of a given type of named entity would entail positive sentiment or negative sentiment.”

It was decided to look more closely into the entities found by NLTK. NLTK found 4 entities, such as “location”, “person”, “organization” and “facility”. The presence of such category for each review was analyzed first. Average review rating was 4.03. That was compared to the average of reviews with specific named entity. Entity presence seemed to correlate with positive sentiments:



1. AVG rating

## Hypothesis 3: “Negative sentiment entail more argumentation.”

Presence of argumentation words is interesting area of NLP. The hypothesis was that negative sentiment entails more argumentation. For this analysis we considered the list of words that induce explanation, such as “because”, “for the purpose”, “since”. The analysis revealed that explanatory words entail negative sentiment. This was proven true with values by average explanatory expression in positive and negative review.

Result for positive review: 0.15483870967741936.

Result for negative review: 0.3677363399826539.

## Hypothesis 4a: “Badly written reviews are likely to be included in ambiguous class”

Ambiguity of each review was determined from VADER result. If the result had significant deviation from the users own rating, it was considered ambiguous. In the dataset, there were very few ambiguous reviews. As a discrepancy value, we used those that deviated by 0.25 or more. Badly written reviews were determined by the percentage of known words in the review. For this, WordNet was used.

Average known words per category showed that percentage for ambiguous was almost same as for unambiguous, both result to 65.6%. As no difference in percentage was found based on our analysis with WordNet, thus, this is false.

## Hypothesis 4b: “Ambiguous reviews are shorter”

Hypothesis 4b claims that ambiguous reviews are shorter than unambiguous. This is easily testable by counting the words in reviews. On average, the total words of the reviews are as per following:

Result Ambiguous: 50.37847222222222

Result Unambiguous: 68.77193163097199

Ambiguous reviews seem to be on average 27% shorter than the unambiguous ones. This hypothesis is proven true.

## Hypothesis 4c: “Ambiguous reviews have bad readability”

To determine if the ambiguous reviews have bad readability, project compared the ambiguousness to Automated Readability Index. Within the dataset, around 97 % of reviews were unambiguous. Out of those, the average ARI was 7.63. For the ambiguous reviews, the ARI was 7.13. This hypothesis was tested to be false.

# Overall Discussion

During the project, NLP pipeline application was developed to analyze the data for sentiment analysis of hotel reviews. The pipeline and the performed steps were mandated in the project description, which was followed by the project team. The data, that was available for the analysis, was examined first before choosing which one to use. The decision was to use Datafiniti hotel reviews dataset.

The process of creating the application, running through different tasks, learning about correlations can be seen as

You should elaborate on the overall findings of your project and potential link with state of art raised in the introduction section. Report also any subsequent work that might be needed to perform specification but cannot be done in the current task

Report any comparison results, if available

Report what you think you have performed as novel and worth pursuing.

# Conclusion

Main conclusion of your work, difficulty of the tasks performed, skills gained and potential ways forward

Use either SI (MKS) or CGS as primary units. (SI units are encouraged.) English units may be used as secondary units (in parentheses). An exception would be the use of English units as identifiers in trade, such as “3.5-inch disk drive”.

* Avoid combining SI and CGS units, such as current in amperes and magnetic field in oersteds. This often leads to confusion because equations do not balance dimensionally. If you must use mixed units, clearly state the units for each quantity that you use in an equation.
* Do not mix complete spellings and abbreviations of units: “Wb/m2” or “webers per square meter”, not “webers/m2”. Spell out units when they appear in text: “. . . a few henries”, not “. . . a few H”.
* Use a zero before decimal points: “0.25”, not “.25”. Use “cm3”, not “cc”. (*bullet list*)

## Equations

The equations are an exception to the prescribed specifications of this template. You will need to determine whether or not your equation should be typed using either the Times New Roman or the Symbol font (please no other font). To create multileveled equations, it may be necessary to treat the equation as a graphic and insert it into the text after your paper is styled.

Number equations consecutively. Equation numbers, within parentheses, are to position flush right, as in (1), using a right tab stop. To make your equations more compact, you may use the solidus ( / ), the exp function, or appropriate exponents. Italicize Roman symbols for quantities and variables, but not Greek symbols. Use a long dash rather than a hyphen for a minus sign. Punctuate equations with commas or periods when they are part of a sentence, as in:

*a**b* 

Note that the equation is centered using a center tab stop. Be sure that the symbols in your equation have been defined before or immediately following the equation. Use “(1)”, not “Eq. (1)” or “equation (1)”, except at the beginning of a sentence: “Equation (1) is . . .”

## Some Common Mistakes

* The word “data” is plural, not singular.
* The subscript for the permeability of vacuum **0, and other common scientific constants, is zero with subscript formatting, not a lowercase letter “o”.
* In American English, commas, semicolons, periods, question and exclamation marks are located within quotation marks only when a complete thought or name is cited, such as a title or full quotation. When quotation marks are used, instead of a bold or italic typeface, to highlight a word or phrase, punctuation should appear outside of the quotation marks. A parenthetical phrase or statement at the end of a
* sentence is punctuated outside of the closing parenthesis (like this). (A parenthetical sentence is punctuated within the parentheses.)
* A graph within a graph is an “inset”, not an “insert”. The word alternatively is preferred to the word “alternately” (unless you really mean something that alternates).
* Do not use the word “essentially” to mean “approximately” or “effectively”.
* In your paper title, if the words “that uses” can accurately replace the word “using”, capitalize the “u”; if not, keep using lower-cased.
* Be aware of the different meanings of the homophones “affect” and “effect”, “complement” and “compliment”, “discreet” and “discrete”, “principal” and “principle”.
* Do not confuse “imply” and “infer”.
* The prefix “non” is not a word; it should be joined to the word it modifies, usually without a hyphen.
* There is no period after the “et” in the Latin abbreviation “et al.”.
* The abbreviation “i.e.” means “that is”, and the abbreviation “e.g.” means “for example”.

An excellent style manual for science writers is [7].

# Using the Template

After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command, and use the naming convention prescribed by your conference for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper; use the scroll down window on the left of the MS Word Formatting toolbar.

## Authors and Affiliations

**The template is designed for, but not limited to, six authors.** A minimum of one author is required for all conference articles. Author names should be listed starting from left to right and then moving down to the next line. This is the author sequence that will be used in future citations and by indexing services. Names should not be listed in columns nor group by affiliation. Please keep your affiliations as succinct as possible (for example, do not differentiate among departments of the same organization).

### For papers with more than six authors: Add author names horizontally, moving to a third row if needed for more than 8 authors.

### For papers with less than six authors: To change the default, adjust the template as follows.

#### Selection: Highlight all author and affiliation lines.

#### Change number of columns: Select the Columns icon from the MS Word Standard toolbar and then select the correct number of columns from the selection palette.

#### Deletion: Delete the author and affiliation lines for the extra authors.

## Identify the Headings

Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

Component heads identify the different components of your paper and are not topically subordinate to each other. Examples include Acknowledgments and References and, for these, the correct style to use is “Heading 5”. Use “figure caption” for your Figure captions, and “table head” for your table title. Run-in heads, such as “Abstract”, will require you to apply a style (in this case, italic) in addition to the style provided by the drop down menu to differentiate the head from the text.

Text heads organize the topics on a relational, hierarchical basis. For example, the paper title is the primary text head because all subsequent material relates and elaborates on this one topic. If there are two or more sub-topics, the next level head (uppercase Roman numerals) should be used and, conversely, if there are not at least two sub-topics, then no subheads should be introduced. Styles named “Heading 1”, “Heading 2”, “Heading 3”, and “Heading 4” are prescribed.

## Figures and Tables

#### Positioning Figures and Tables: Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation “Fig. 1”, even at the beginning of a sentence.

1. Table Type Styles

| Table Head | Table Column Head | | |
| --- | --- | --- | --- |
| Table column subhead | Subhead | Subhead |
| copy | More table copya |  |  |

1. Sample of a Table footnote. (*Table footnote*)
2. Example of a figure caption. (*figure caption*)

Figure Labels: Use 8 point Times New Roman for Figure labels. Use words rather than symbols or abbreviations when writing Figure axis labels to avoid confusing the reader. As an example, write the quantity “Magnetization”, or “Magnetization, M”, not just “M”. If including units in the label, present them within parentheses. Do not label axes only with units. In the example, write “Magnetization (A/m)” or “Magnetization {A[m(1)]}”, not just “A/m”. Do not label axes with a ratio of quantities and units. For example, write “Temperature (K)”, not “Temperature/K”.

##### Acknowledgment *(Heading 5)*

The preferred spelling of the word “acknowledgment” in America is without an “e” after the “g”. Avoid the stilted expression “one of us (R. B. G.) thanks ...”. Instead, try “R. B. G. thanks...”. Put sponsor acknowledgments in the unnumbered footnote on the first page.

##### References

The template will number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use “Ref. [3]” or “reference [3]” except at the beginning of a sentence: “Reference [3] was the first ...”

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the abstract or reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors’ names; do not use “et al.”. Papers that have not been published, even if they have been submitted for publication, should be cited as “unpublished” [4]. Papers that have been accepted for publication should be cited as “in press” [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

1. G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. *(references)*
2. J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.
3. I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.
4. K. Elissa, “Title of paper if known,” unpublished.
5. R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.
6. Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].
7. M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.

**IEEE conference templates contain guidance text for composing and formatting conference papers. Please ensure that all template text is removed from your conference paper prior to submission to the conference. Failure to remove template text from your paper may result in your paper not being published.**

We suggest that you use a text box to insert a graphic (which is ideally a 300 dpi TIFF or EPS file, with all fonts embedded) because, in an MSW document, this method is somewhat more stable than directly inserting a picture.

To have non-visible rules on your frame, use the MSWord “Format” pull-down menu, select Text Box > Colors and Lines to choose No Fill and No Line.