CHALMERS
UNIVERSITY OF TECHNOLOGY

UNIVERSITY OF GOTHENBURG

# Comparative Evaluation of Sentiment Analysis Methods

Master's thesis in Computer science and engineering

KIERON HAYES

Department of Computer Science and Engineering
CHALMERS UNIVERSITY OF TECHNOLOGY
UNIVERSITY OF GOTHENBURG
Gothenburg, Sweden 2022

Master's thesis 2022

# Comparative Evaluation of Sentiment Analysis Methods

Kieron Hayes

UNIVERSITY OF GOTHENBURG

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Computer Science and Engineering
Chalmers University of Technology
University of Gothenburg
Gothenburg, Sweden 2022

Comparative Evaluation of Sentiment Analysis Methods
KIERON HAYES

Supervisor: Aleksandrs Berdicevskis, Department of Swedish, multilingualism, language technology
Examiner: Richard Johansson, Department of Computer Science and Engineering

Master's Thesis 2022
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Typeset in LaTeX
Gothenburg, Sweden 2022

Comparative Evaluation of Sentiment Analysis Methods
Kieron Hayes
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg

# Abstract

Sentiment analysis is an on-going field of research within the realm of Natural Language Processing, in which we wish to accurately assess the sentiment of an author on a given topic. Within this thesis project I construct a rule-based system for sentiment analysis specific to the domain of hard rock and heavy metal music album reviews. I then compare it to the performance of other approaches to the task, such as the use of a neural network, and analyse the strengths and weaknesses of these differing approaches. Ultimately the neural network, with sufficient training, produces the best results for this task, and I go on to outline possible improvements that could be made to the rule-based system in further efforts to maximise its potential.

Keywords: Sentiment analysis, BERT, Rule-based, Domain-specific, SentiWordNet

# Acknowledgements

I would like to sincerely thank my supervisor Aleksandrs Berdicevskis for taking on my thesis project in the first place, and for always being able to rely on him for guidance and invaluable knowledge, as well as my examiner Richard Johansson for inspiring me to work within the field of Natural Language Processing through his excellent teaching. I would also like to thank Evelina Strauss and Usama Safdar, my opponents for this thesis, for the inspiration of their own hard work, and fellow data science student Atefeh Aminmoghaddam for being such a wonderful groupmate on many courses during my first year. Finally, I would like to thank my partner and my parents, as without their constant support and encouragement I could never have made it this far.

Kieron Hayes, Gothenburg, June 2022

# Contents

# List of Figures

# List of Tables

# 1
# Introduction

This chapter introduces the field of sentiment analysis and the goals of this thesis project.

## 1.1 Background

Language and communication have been vital parts of human societies for as long as those societies have existed, and just as languages themselves have changed over time, so too have our needs regarding them, as our world develops. The invention of modern computers also brought with it the creation of code, as a means to effectively communicate with these machines, to convey our instructions to them. The famous Turing Test (originally termed *The Imitation Game*) measures a computer's intelligence by its capability to communicate with a human being without that person being able to tell the difference [8]. The internet acts as both a method for and repository of a vast wealth of human communications.

This connectivity between human communication and machines has naturally given rise to a discipline devoted to its study and refinement. Natural Language Processing (NLP) is the application of machine learning tools to examine natural human language (written or verbal) and make use of it in various ways, be this for the purposes of translation, discerning patterns across a large corpus, synthesising natural language from a computer, or extracting people's opinions on a particular topic [4]. The focus on accurately detecting opinions specifically is referred to as sentiment analysis.

The growth of the internet, both in terms of its size and relevance in modern life, has simultaneously provided a vast source of data and increased the need for powerful, efficient NLP implementations. The potential wealth of available data for analysis continues to grow daily: the internet as a whole is estimated to currently be some 5 million terabytes worth of data [9]. Social media platforms like Facebook, Twitter and Reddit allow users to express a constant stream of thoughts and opinions, while many websites act as repositories for written reviews, from online stores like Amazon, to widely utilised sources of critical and general opinion such as Rotten Tomatoes or IMDB.

As the scale of the data available grows, so too do the needs for special methods to handle it. Those seeking information or opinions on a particular topic or product will not be short of available data, but may instead be hampered by the need for a tool to sift through this data and extract meaningful, useful results.

Codified scores such as star ratings can certainly help in this regard, allowing users to quickly and easily see how well a product has been received. But even this may not always be sufficient, especially regarding more subjective works whose nature is less functional and more about enjoyment. While a critic's assigned score to a movie or music album can be useful at a shallow level, the deeper rationale behind that rating may be more useful. Elements that may be strongly disliked by one individual may be taken in a more positive light to another. As such, there is great potential utility in a tool's capability to highlight the most important points (both positive and negative) of a subjective review.

## 1.2 Aims

As individuals, we are capable of reading a text (or listening to speech) and detecting opinions expressed, and as will be detailed further within Chapter 2 (Theory), there are different approaches to achieving this with machines. The core aim of this project is to examine one particular method, that of a rule-based lexicon approach, and to compare it to other potential methods for achieving the same end result.

As such, the most significant single task will be the construction of a system that can implement such an approach. This will begin with close manual examination of a number of sample reviews, to attempt to identify the patterns within the language that could then form the basis of the system's ruleset. In other words, how can we reliably identify the sentiment of the texts when reading them, and in turn, how can we convert this into a rule or set of rules for the coded system?

The resulting system will need to make predictions for the overall rating of a review (representing the reviewer's feelings about it as a number). Thus, we must aim to craft a system that can produce predicted ratings that are as accurate as possible. The precise details of how this system is evaluated will be covered later within the report.

In addition to making these rating predictions, I also aim to have this system output sample sentences from the review as rationales justifying its decision. This will be done by assessing each sentence for its sentiment polarity weight (i.e. how much did it contribute to the overall prediction), then returning the sentences with the greatest weight. This will point to the sentences that had the greatest impact, which could in turn be taken as something of a summary of the review's most important points.

As well as producing an accurate rule-based system, this project also aims to evaluate this in comparison to other methods, to examine how these differing approaches may perform compared to one another. Does the work of establishing a set of rules provide enough of a benefit when compared to a much simpler system that does not pay attention to linguistic elements beyond the sentiment of each individual word? Can a rule-based system match or even exceed the performance of a pre-trained neural network?

Lastly, I intend to examine whether the systems could be used to complement one another, one helping to overcome the shortcomings of another. There is a possibility that the aforementioned rationale outputs of the rule-based system could be utilised as a form of diagnostic tool for a neural network system. By removing these highlighted sentences from the inputs and then examining the change in the network's outputs, we may gain some greater insight into the network's methods and how it makes the predictions it does.

Given the above, the research questions for this project can be considered as the following:

1. How does a rule-based approach to sentiment analysis compare to methods such as a neural network or simple count-based baseline?

2. Do domain-specific modifications improve the performance of sentiment analysis within this domain?

3. Can a rule-based system be made to output strongly polar sentences as rationales for its decision, and do these accurately represent some of the most important points of a review?

4. Can these rationale outputs be used like a diagnostic tool for a neural network to help provide greater insight into its process?

## 1.3 Limitations

While the work done over the course of this thesis project naturally aims to be as comprehensive as is feasible, naturally as the work of one individual over a number of months it comes with certain limitations of scope.

The work done within this project will be domain specific by design. It will utilise a corpus comprised of music album reviews, all drawn from websites with a focus on hard rock and heavy metal styles of music. Different domains often require unique approaches or modifications to accommodate their own traits, and so this should not be taken as a hindrance in itself, but it is important to note this domain specificity as a key element of the project. The sole use of online publications over print publications, while done for practical reasons, could also be a further restriction in this regard. The corpus will also consist solely of English language entries (certain elements such as album or song titles may include use of other languages, but these will not be factored in to the rules of the system specifically).

The primary point of focus for investigation is the rule-based system. The identification of the linguistic patterns its rules will be based on, and the construction of a system to implement these rules, will take up a significant portion of the available time. While other systems will be utilised for comparison purposes, substantially less time will be spent on the development and refinement of these compared to the rule-based system. In the case of the neural network, an existing pre-trained model (Bidirectional Encoder Representations from Transformers, BERT [33]) will be utilised, rather than attempting to build a custom network from the ground up. This is intended to investigate a comparison between a carefully constructed rule-based lexicon and an *off-the-shelf* neural network as might be realistically employed for such a task today.

The rule-based system will also make use of a number of existing tools for different parts of its functionality. For example, it will utilise an existing parsing tool for identifying word dependency relationships within a sentence, as well as an existing method for disambiguating words and determining their exact meaning. These components will be covered in more detail in Chapter 3 (Methodology). While some of these may be modified in certain ways (such as modifying SentiWordNet 3.0 to make it more appropriate for this specific domain), there is only so much that can realistically be done, and these components will naturally come with their own imperfections.

In addition, there are a number of known challenges within the field of sentiment analysis which this project does not attempt to address, as to do so would be far too ambitious given the time and resources available. The use of irony, for example, can be difficult for systems to accurately identify [15]. To properly tackle this issue could make up a thesis project all of its own, and I will not be attempting to resolve it here.

Lastly, while the corpus includes existing labels for overall album rating, it could be useful to annotate it further; for example, by adding labels for sentiment expressions about specific elements of the album (see the *aspects* detailed within Section 3). However, there is not sufficient time to perform manual annotation of this nature across the full corpus. This can and will be done on a smaller sub-set during the evaluation process though.

## 1.4 Report outline

In this first chapter, I have given an outline of the problem space and what this project is aiming to achieve.

- In Chapter 2, I will further explore the topics of NLP and sentiment analysis, including common approaches within the field, and give some additional details on existing tools that will be utilised within this project.

- In Chapter 3, I will present the detailed methodology of the project. This will include the construction of a corpus specifically for the project, how the core rule-based system was built and expanded upon, the creation of two other systems for comparative evaluation, and hyperparameter tuning.

- In Chapter 4, I will present the results of running all systems through the full test set of the corpus, as well as results from investigated other specific points of interest.

- In Chapter 5, the results from the previous section will be analysed. This will include a straightforward comparison of numerical metric results, and a more qualitative assessment of their performances, with analysis of error results to provide insight into the systems' shortcomings.

- In Chapter 6, I will discuss what over-arching conclusions I believe can be drawn about the project based on its results. This section will also include discussion of problems or limitations that were identified about the project and hypothetical future avenues for expansion or refinement.

# 2
# Theory

This chapter further explores the topics of NLP and sentiment analysis, including common approaches within the field, and gives some additional details on existing tools that will be utilised within this project.

## 2.1 Existing work

As outlined in the previous section, sentiment analysis is the use of NLP methods to assess inputs (typically text, though it could also include speech) and determine the sentiments expressed therein, if any. Exactly how this is implemented, and how its success is measured, can vary substantially. Typically, sentiment is measured as positive or negative (or neutral), and this could be treated categorically (treating *positive* and *negative* as distinct category labels) or numerically (measuring sentiment along a numerical scale).

Textual inputs will naturally require certain pre-processing steps in order for coded systems to operate on them as intended, and while the exact details may vary depending on the process used, there are some common steps. These include:

- **Tokenisation**: the components of a sentence are broken up into distinct tokens for the system to handle (which can include elements beyond words, such as punctuation)

- **Part-of-speech tagging**: tokens are tagged with markers indicating what role they serve within this piece of text/speech, e.g. identifying whether a word is a noun, adjective, verb, etc.

- **Lemmatisation**: words are transformed into their base form (so that *dog* and *dogs* will be recognised as having the same basic meaning). It is worth noting that in the case of lemmatisation, the original word form can still be retained, so this data is not lost, but a lemmatised *base form* is available for operations where variations of a core word should be treated the same way.

Such preprocessing may also involve reducing words to lowercase form, though again the original form of the word may be retained for operations where the capitalisation is valuable data.

*Parsing* is also a common practice within NLP. This refers to the identification of the syntactic structure of a piece of text, done in accordance with the grammatical rules of the language in question. Within this, links can be established between words, creating a dependency tree, as illustrated in Figure 2.1. This parsing process will be vital in the rule-based system created within this thesis, and its function and outputs will be described in greater detail later.

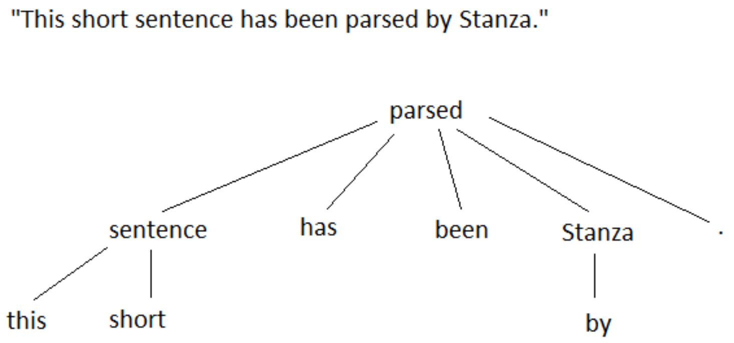


**Figure 2.1:** Example of word dependency tree created through parsing.

Existing work often identifies differing levels of sentiment analysis, including document-level and sentence-level [4]. In *document-level work*, whole documents are classified as positive or negative. A review of a product, for example, might include individual sections that express different kinds of sentiment, but the full document will be given a single label or score. In 2002, Pang, Lee & Vaithyanathan took several models traditionally used for document topic classification and attempted to apply them to sentiment classification of these documents instead [6]. Their methodology involved several concepts that would be recurring themes in other works in the field, such as designating negation words (e.g. *not*) which would flip the polarity of sentiments, and using part-of-speech tagging (in which a word is tagged to identify what role it fulfils within a sentence, e.g. whether it is a noun, adjective, verb, etc). They concluded that while sentiment classification could be done, it appeared more difficult than topic classification, and the models used struggled to match performance in the former as in the latter. These conclusions, combined with the methodological points specific to this domain, highlight how special approaches are needed within the field of sentiment analysis to achieve good results.

At *sentence-level*, individual sentences are assigned polarity (which can also include being neutral), though in some instances the analysis may go deeper still, splitting a sentence up into multiple smaller units (e.g. phrases). This allows for a document to express many different opinions on a single (or even multiple) topics. In 2003, Nasukawa and Yi [3] conducted a study focusing at this level of analysis, in which they constructed a lexicon of terms identifying both sentiment and how these words might interact with one another. Seeming to take a step closer to actual human comprehension of language, this paper experiments with taking a deeper look at the mechanics of the language (in this case English), discerning how words might pass sentiment on to different topics within a single sentence. This approach, focusing on encoding the syntactic meaning of words, has been utilised further in papers such as *Recognising Contextual Polarity in Phrase-Level Sentiment Analysis* [5].

While not strictly a contrasting method, *aspect-based* analysis is another important approach to sentiment analysis. Rather than assigning sentiment to a document or sentence as a whole, in this approach the system aims to identify different aspects of an entity and judge the sentiment expressed about those aspects separately. For example, in a movie review there could be a sentence reading:

> *The direction was outstanding, and the acting was fantastic, even if the soundtrack fell flat.*

The sentence seems to express three sentiments: positivity about the direction and acting, but negativity about the soundtrack. Depending on exactly how the sentiment is measured and recorded, either the full sentence may be classified as positive (discarding the sentiment expressed about the soundtrack) or the sentence may impart a mixture of both positivity and negativity. But in an aspect-based approach, these sentiments would be assigned separately to different aspects of the movie: the scores of the direction/director and acting/actors would increase, while the score of the soundtrack would decrease.

Singh et al. investigated this very topic in 2013. They identified specific aspects of movies within reviews, then sought to assign scores to those specific elements [2]. These aspects were selected by searching through movie reviews, film magazines, and movie award categories. Aspect vectors could then be created to represent the words that express that aspect (e.g. *screenplay*). Reviews were then parsed sentence-by-sentence to find opinions expressed about these aspects, and SWN was used to determine the polarity.

The resulting outputs acted as a breakdown of how these different aspects, such as editing, cinematography, music, acting performance, etc, were rated according to reviewers. An example of an output breakdown can be seen in Figure 2.2, where we see a record of the overall positivity and negativity expressed across a review document about different aspects of the film *No Problem.*

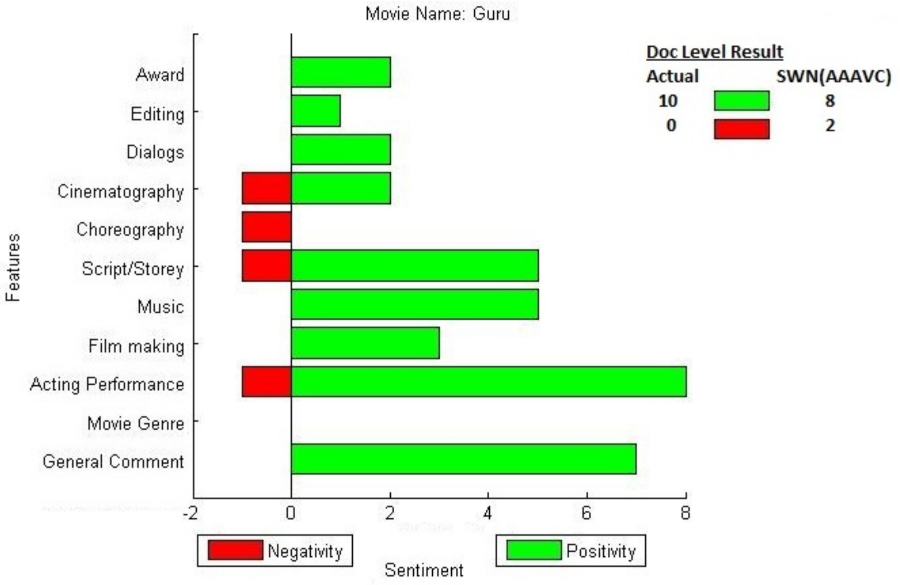


**Figure 2.2:** Aspect-based assessment of the movie No Problem from Sentiment Analysis of Movie Reviews [2].

A further division within the field exists between rules- or lexicon-based approaches and those who treat sentiment analysis as a text classification task utilising machine-learning models [14].

### 2.1.1 Rule-based approaches

In a rule-based approach, specific rules are codified to indicate the meaning of the text being processed. Typically, a lexicon will be used to record the details of these rules, and a system will be constructed that can take in text, analyse it with regards to these rules, and assess sentiment as a result. A simple example of this could be a dictionary of words and a binary classification of their sentiment (positive or negative). Every word within a given text could thus be checked against this dictionary, and the full document could be classified based on how many positive or negative terms are detected. In the example below, positive words are marked in green, and negative in red:

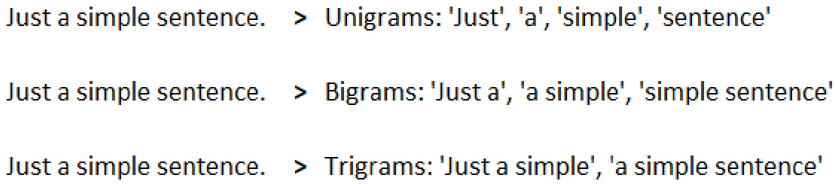
> *The direction was outstanding, and the acting was fantastic, even if the soundtrack fell flat.*

Naturally, the true sentiment of a statement often depends on more than simply the meaning of individual words in isolation: the way words are put together, the context of neighbouring words, specific combinations of words or phrases, these can all influence the true meaning that we take from text or speech. In turn, efforts have been made to incorporate wider elements of language into lexicons, such as independent clauses [16], conjuncts [17], and the influence of context within phrases [5].

### 2.1.2 Machine learning

Machine learning tools are commonly used for NLP related tasks, including sentiment analysis. With sentiment analysis, this is often addressed as a supervised learning classification task: a model is trained on sample texts labelled with the ground truth label (e.g. positive or negative), and learns to make an association between the contents of the texts and the desired output. Support Vector Machine classifiers are commonly used for these tasks [14]. These will encode text features as vectors, with semantic similarities between words represented through the features of these vectors.

While such tasks may take a simple bag-of-words approach (where words are treated separately), they may often make use of other techniques to better capture the intricacies of language, in which the context of a word can significantly impact its meaning. One such method is the use of n-grams, in which sequences of words (of length $n$) are examined rather than just single words alone, allowing a system to capture more contextual information about a word and its neighbours. Different values of n-grams were tested by Pang, Lee & Vaithyanathan, who found the use of bigrams did not significantly alter their results from the use of unigrams [6].



Just a simple sentence. > Unigrams: 'Just', 'a', 'simple', 'sentence'

Just a simple sentence. > Bigrams: 'Just a', 'a simple', 'simple sentence'

Just a simple sentence. > Trigrams: 'Just a simple', 'a simple sentence'

**Figure 2.3:** Illustration of n-grams.

Neural networks are also commonly used within the field of NLP. A network is constructed with multiple layers that transform their given inputs, then pass this forward to the next layer. These layers will typically pass over input in a specific way (e.g. utilising n-grams of words as detailed above), then pass on the results from this, in this way seeking out patterns within the data. At the end of a network is a chosen model type which will take the final outputs of the previous layers and produce a final result, depending on the desired output type (e.g. a predicted class for a classification task).

For a sentiment analysis task, a network may be trained on a number of sentiment documents labelled as positive or negative, learning to draw connections between patterns within the text (e.g. the presence of certain words or phrases) and the label. Typically, the input data will be represented as vectors, and these representations are encoded within the network's layers as embeddings. These embeddings represent the words as machine-readable real-valued vectors. Once the network has been trained, it can then take in a new document and search for these indicative patterns to make a new prediction.

As with other machine learning tasks, embeddings have become a key component of neural networks for NLP, as these transform words into mathematical vectors, with the distances between these vectors assumed to represent real semantic distinctions or similarities between words [31].

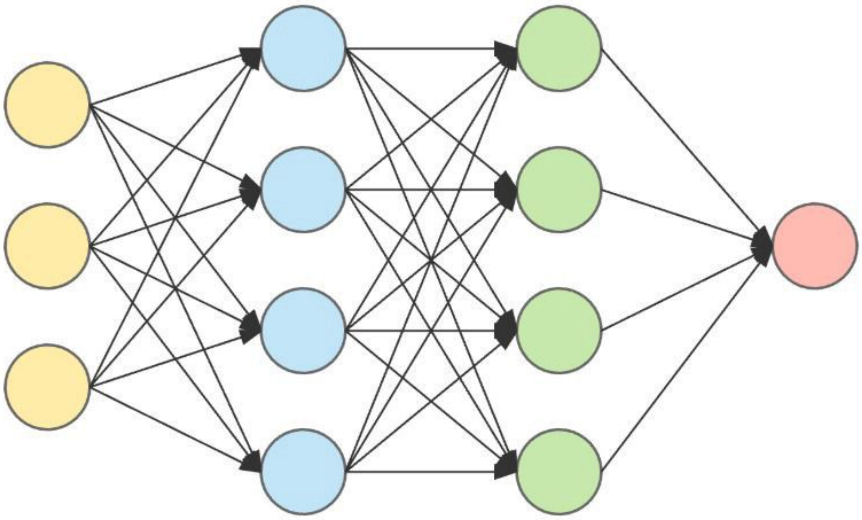


**Figure 2.4:** Illustration of the layers of a neural network passing to one another [32].

Modern approaches to sentiment analysis of text have tended to focus on deep learning and complex neural networks, and these have produced good results [11]. However, these results can be difficult to examine and fully interpret, especially in larger models with many hidden layers and connections [12]. Thus, while these models may produce accurate results, it can be difficult to truly understand how each result was reached.

#### 2.1.2.1 BERT

BERT (Bidirectional Encoder Representations from Transformers) is a language representation model. Rather than the traditional method of reading input text sequentially left-to-right or right-to-left, BERT trains transformers by reading in both directions simultaneously and building connections between different elements [33]. This enables it to more accurately assess the true meaning of a word, taking context into account and better tackling ambiguity: where older models may map a word to a representative vector, this alone does not factor in the contextual meaning of the word, and even when attention is paid to the sentence as a whole, reading in one particular direction may result in biases. By reading in this bidirectional manner, BERT can more clearly see how a word within a sentence is impacted by all others.

Typically BERT will be loaded as a pre-trained model, having been trained on unlabelled data (English Wikipedia for example) across different tasks, then fine-tuned on labelled data for a specific task or tasks. The transformer output from

BERT can be passed into a final layer to convert that output into a desired prediction (e.g. a layer that converts it into a class for a classification problem).

## 2.2 Future avenues

In their survey paper *Evaluative Language Beyond Bags of Words*, Benamara, Taboada and Mathieu outline what they believe to be a number of different key avenues for future work in sentiment analysis [21]. These include:

- **Intent detection** - Determining the intent of the author, and how this might influence sentiment expressed.

- **Implicit evaluation** - Detecting sentiment where it is not explicitly stated, but implied, e.g. through association ("This book is as good as Lord of the Rings" only conveys sentiment if you are aware of general opinion of Lord of the Rings).

- **Extra-linguistic information** - Details that may be relevant to sentiment but lie beyond the text itself, such as the demographics of the author or their communication medium.

- **Figurative language** - Metaphors, idioms, oxymorons, irony, etc.

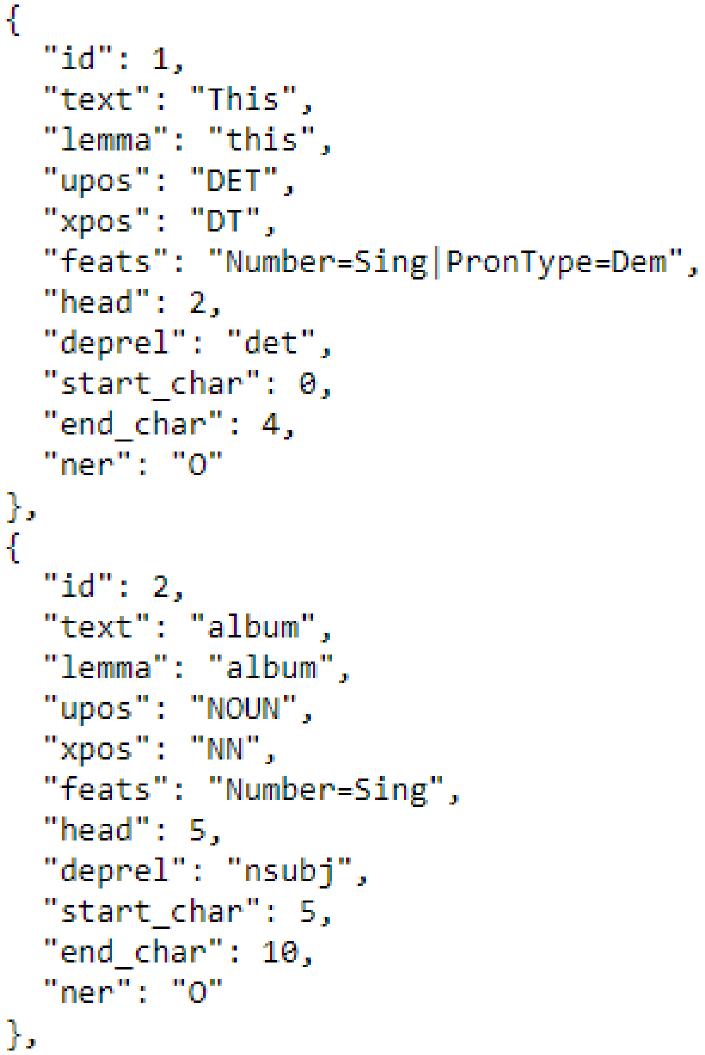
- **Discourse analysis** - In which a sentiment document is viewed as a piece of discourse between writer and audience, with a certain flow and structure to it that can be useful in analysing sentiment expressed therein. This sort of analysis may be approached in either a top-down or bottom-up manner. The bottom-up approach focuses on Elementary Discourse Units (EDUs) as a way to categorise units of discourse, and examines their types and the relationships between them. The top-down approach examines higher-level patterns within a text, looking at the larger-than-sentence segments and how they interact with one another [21].

## 2.3 Existing tools

Much work has been done within the realm of sentiment analysis, and many tools and models have been created and made available. These can act as sources of inspiration or could be directly employed as components within a wider project. Several that will be of particular relevance to this thesis project are introduced below.

### 2.3.1 Stanza

Stanza is a NLP toolkit package for the Python programming language [25], pre-trained on the Universal Dependencies framework [39]. It includes a pipeline in which an input sentence may be processed by a number of tools in order to transform the text and extract important information for use in NLP projects. This includes lemmatisation of words, applying part-of-speech tags, identifying morphological features of a word (see the example below), and a syntactic structure dependency parse which links words together in a dependency tree. In this way, the syntactic links between words can be established and operated upon within Python.



```
{
  "id": 1,
  "text": "This",
  "lemma": "this",
  "upos": "DET",
  "xpos": "DT",
  "feats": "Number=Sing|PronType=Dem",
  "head": 2,
  "deprel": "det",
  "start_char": 0,
  "end_char": 4,
  "ner": "O"
},
{
  "id": 2,
  "text": "album",
  "lemma": "album",
  "upos": "NOUN",
  "xpos": "NN",
  "feats": "Number=Sing",
  "head": 5,
  "deprel": "nsubj",
  "start_char": 5,
  "end_char": 10,
  "ner": "O"
},
```

**Figure 2.5:** Examples of word details returned by Stanza pipeline.

### 2.3.2 SentiWordNet

A development of the WordNet lexical database, SentiWordNet (SWN) assigns polarity markers to words, indicating their sentiment value [10]. This is done on a pair of scales, PN polarity (representing their subjectivity, positive vs. negative) and SO polarity (representing their objectivity, subjective vs. objective). Each word has three values, positive, negative and objectivity, which combined always sum to 1. In its most recent version, 3.0, SWN is based on WordNet 3.0 [20], which contains 155,287 unique words [26].

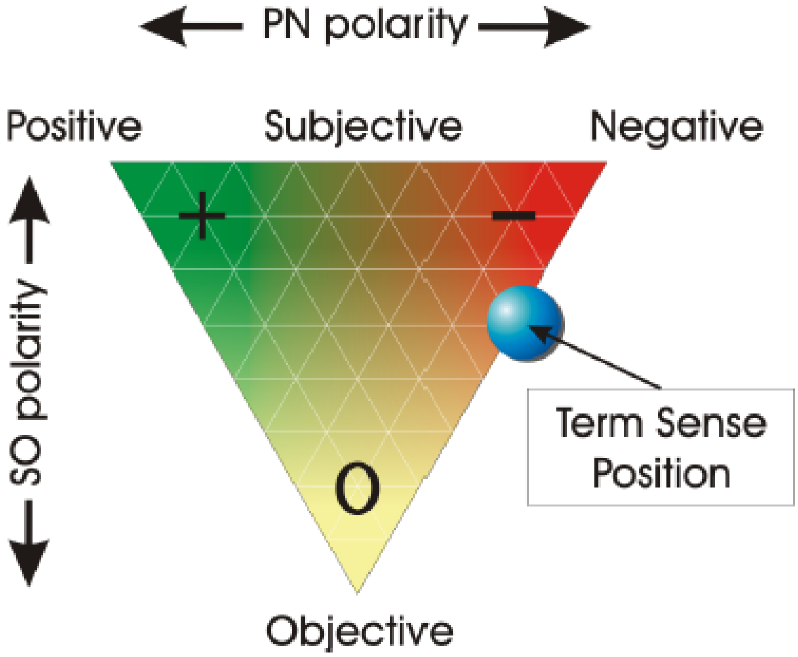


**Figure 2.6:** The graphical representation adopted by SENTI-WORDNET for representing the opinion-related properties of a term sense [10].

### 2.3.3 NLTK Word Sense Disambiguation

NLTK (Natural Language Toolkit) is a toolset for the Python programming language [27]. It contains a wide range of NLP-related functionalities, but of particular interest here is its capability for Word Sense Disambiguation (WSD). This refers to the fact that a word may have more than one meaning (e.g. *wave* could be used as a verb, *to wave one's hand*, or as a noun, referring to a wave on a body of water). Thus, any system assessing words in a document needs a way to make a decision about which sense a word is used in. NLTK includes one such method.

### 2.3.4 SO-CAL

SO-CAL (Semantic Orientation CALculator) is an existing sentiment analysis tool [14]. Of particular note for this thesis project is its handling of intensifiers and negation terms. When intensifiers are found within text they apply a percentage modifier to relevant terms (so *very good* would increase the positive impact of *good* by an additional 25%, while *slightly good* would decrease it by 50%). Negation is handled in a similar way, as the authors make the observation that a negation term does not always reverse or totally invalidate a sentiment (this method being known as switch negation [28]). *Not good* is not truly the polar opposite of *good*, for example. Instead, they employ a method of shift negation that modifies polarity by a fixed amount.

# 3
# Methodology

This chapter presents the detailed methodology of the project. This includes the construction of a corpus specifically for the project, how the core rule-based system was built and expanded upon, the creation of two other systems for comparative evaluation, and hyperparameter tuning.

## 3.1 Corpus

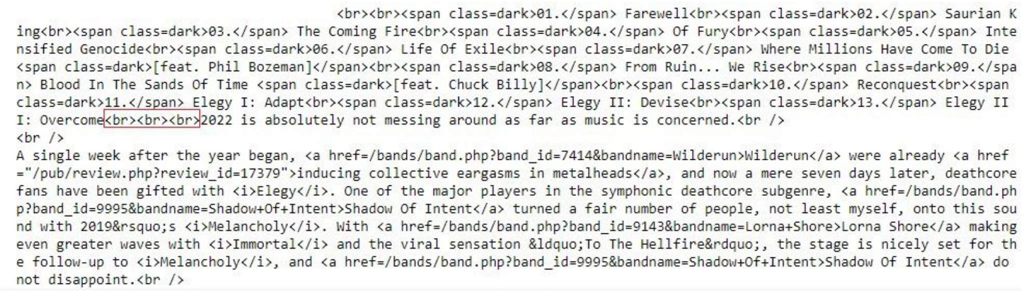
The necessary first step on the project was the creation of the corpus. The core system constructed across this thesis project would be a rule-based system, and as such it would not require training as a machine-learning model would. However, close analysis of corpus examples would be needed in order to identify domain-relevant patterns and establish the ruleset to implement. In turn, once the system was completed, a corpus would be needed to test its performance across a wide number of entries, and for comparison with other systems.

In my personal free time I operate a project called Meta-Metal [24], in which I collect and aggregate review scores for hard rock and heavy metal music albums from about two dozen different sources, both print and digital. As such, I was already very familiar with this particular domain and these sources, and so this was chosen as a specific domain to investigate.

While a large corpus was desirable for the ability to run systems across many different reviews, the entire contents of all of Meta-Metal's sources would be far more than was necessary for this project, and so only some of the available websites would be needed. The print magazines were discarded, as it would have required substantial effort to digitise the contents. Websites with reviews in languages other than English were also set aside, as I felt that attempting to handle other languages would over-complicate the project. Some websites were set up to automatically deny scraping attempts, so these too were set aside, as acquiring the necessary data from them manually would be extremely time-consuming, and more than enough potential websites were available which accepted scraping attempts.

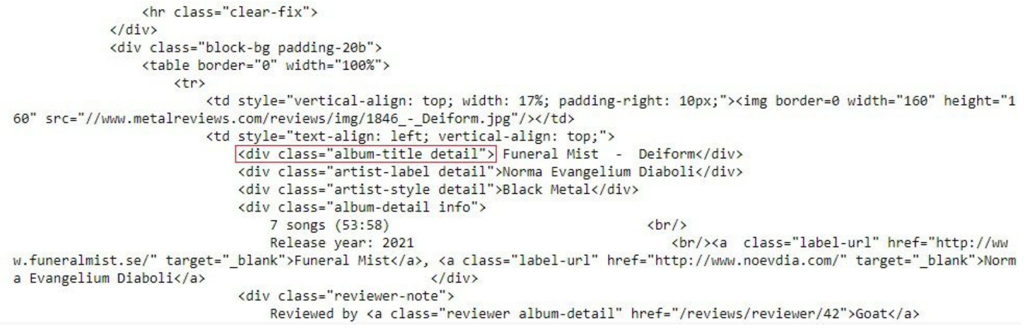
It then became a question of which websites would be the easiest to scrape for the data needed. For this project, the following elements were needed for each review: a title (in the format [*band name*] - [*album title*]), the score assigned to the album by the reviewer and the review text itself. Unique ID numbers and the website source would also be added as additional fields, but these would not need to be scraped from the websites themselves. It should be noted that the unique ID numbers are retained as the corpus is split apart.

Each website has its own particular code architecture. Samples from each site were taken using the *urlopen* scraping method in Python, which takes in a given URL address and returns the raw contents of that page. The returned results were then be examined to search for reliable markers of where the desired data fields started and ended. For example, if a website included a marker, *reviewscore=*, followed by the score given to that album, then a system could be set up to scrape each review page, find the number that follows this marker, and return that as the review score. The easiest sites to scrape would be those with markers that reliably appeared at the start/end of the desired sections of data, and did not appear anywhere else. Some real examples of such markers can be seen in Figures 3.1 and 3.2.



```
                                         <br><br><span class=dark>01.</span> Farewell<br><span class=dark>02.</span> Saurian K
ing<br><span class=dark>03.</span> The Coming Fire<br><span class=dark>04.</span> Of Fury<br><span class=dark>05.</span> Inte
nsified Genocide<br><span class=dark>06.</span> Life Of Exile<br><span class=dark>07.</span> Where Millions Have Come To Die
<span class=dark>[feat. Phil Bozeman]</span><br><span class=dark>08.</span> From Ruin... We Rise<br><span class=dark>09.</spa
n> Blood In The Sands Of Time <span class=dark>[feat. Chuck Billy]</span><br><span class=dark>10.</span> Reconquest<br><span
class=dark>11.</span> Elegy I: Adapt<br><span class=dark>12.</span> Elegy II: Devise<br><span class=dark>13.</span> Elegy II
I: Overcome<br><br><br>2022 is absolutely not messing around as far as music is concerned.<br />
<br />
A single week after the year began, <a href=/bands/band.php?band_id=7414&bandname=Wilderun>Wilderun</a> were already <a href
="/pub/review.php?review_id=17379">inducing collective eargasms in metalheads</a>, and now a mere seven days later, deathcore
fans have been gifted with <i>Elegy</i>. One of the major players in the symphonic deathcore subgenre, <a href=/bands/band.ph
p?band_id=9995&bandname=Shadow+Of+Intent>Shadow Of Intent</a> turned a fair number of people, not least myself, onto this sou
nd with 2019&rsquo;s <i>Melancholy</i>. With <a href=/bands/band.php?band_id=9143&bandname=Lorna+Shore>Lorna Shore</a> making
even greater waves with <i>Immortal</i> and the viral sensation &ldquo;To The Hellfire&rdquo;, the stage is nicely set for th
e follow-up to <i>Melancholy</i>, and <a href=/bands/band.php?band_id=9995&bandname=Shadow+Of+Intent>Shadow Of Intent</a> do
not disappoint.<br />
```

**Figure 3.1:** Example of scraping results from metalstorm.net. Here the text in the red box (<br><br><br>) acts as a unique marker for where the review text begins.



```
            <hr class="clear-fix">
        </div>
        <div class="block-bg padding-20b">
            <table border="0" width="100%">
                <tr>
                    <td style="vertical-align: top; width: 17%; padding-right: 10px;"><img border=0 width="160" height="1
60" src="//www.metalreviews.com/reviews/img/1846_-_Deiform.jpg"/></td>
                    <td style="text-align: left; vertical-align: top;">
                        <div class="album-title detail"> Funeral Mist  -  Deiform</div>
                        <div class="artist-label detail">Norma Evangelium Diaboli</div>
                        <div class="artist-style detail">Black Metal</div>
                        <div class="album-detail info">
                            7 songs (53:58)                         <br/>
                            Release year: 2021                         <br/><a  class="label-url" href="http://ww
w.funeralmist.se/" target="_blank">Funeral Mist</a>, <a class="label-url" href="http://www.noevdia.com/" target="_blank">Norm
a Evangelium Diaboli</a>                        </div>
                        <div class="reviewer-note">
                            Reviewed by <a class="reviewer album-detail" href="/reviews/reviewer/42">Goat</a>
```

**Figure 3.2:** Example of scraping result from metalreviews.com. Here the text in the red box (<div class="album-title detail">) acts as a marker for where the review title begins.

Similarly, a method would be needed in some cases to compile all the review URLs from a given site. Again, this would depend on the set-up on that particular site: in some cases, the URL of each review may end with "review=1234", with the number increasing each time a new review is added to the site. In this case, a system could simply replace the number at the end of the URL and thus increment through every review on the website. In other cases, there may be directory pages organising all a site's reviews (e.g. a page for each letter of the alphabet, by band name), and these directory pages could be scraped for all their review links, then these in turn could be scraped for the data I needed.

Once websites had been chosen to perform scraping on, code was built based around the identified markers in order to extract the necessary data while trying to leave behind as much excess as possible. Further work was done to clean the scraped data and make it suitable for use. This involved:

- Stripping away unneeded characters from certain fields, such as removing "Review of" from the start of the title field. The review rating field also required some tweaking, such as removing "/5" from some entries, or amending decimal commas to decimal points. Entries with ratings in formats that could not be worked with were removed entirely, e.g. N/A.

- Removing formatting markers within the review texts (such as <span> markers indicating particular text fonts or colours).

- Handling characters not recognised by the scraping process, as these would not be encoded properly. This was a case of searching for these markers within the scraped reviews (e.g. *&#926;*), checking which character this corresponds to in the review text on the website of origin, then encoding a large dictionary structure in Python that mapped the code to the intended character. Code could then be run to replace these instances with the appropriate characters. This was often an issue with non-English characters, such as when song titles, album titles or band names utilised Asian, Greek or Cyrillic characters.

All of this culminated in a corpus consisting of a unique index number for each review, a title for it (in the format [*band name*] - [*album title*]), the website source of the review, the rating given by the reviewer, and the review text itself. Four websites were used to construct the corpus: metalcrypt.com [35], metalstorm.net [37], metalunderground.com [38] and metalreviews.com [36], and in total this made up 29,084 reviews (2686 from metalunderground.com, 4872 from metalstorm.net, 8728 from metalreviews.com and 12798 from metalcrypt.com). An example from the corpus can be seen in Table 3.1.

**Table 3.1:** Examples from the corpus

| *[index]* | **ID** | **Source** | **Title** | **Rating** | **Review** |
|---|---|---|---|---|---|
| 0 | 9240 | Metal Crypt | Folkearth - A Nordic Poem | 9.0 | *[review text...]* |
| 1 | 11628 | Metal Crypt | Lordi - Get Heavy | 5.5 | *[review text...]* |
| 2 | 24577 | Metal Reviews | Exciter - Thrash Speed Burn | 7.0 | *[review text...]* |
| 3 | 7697 | Metal Crypt | Dawn of Tears - Act III: The Dying Eve | 7.0 | *[review text...]* |
| 4 | 1313 | Metal Storm | Theories - Regression | 8.2 | *[review text...]* |

While the corpus itself cannot be shared (as all reviewers would need to give their consent), the script utilised for this scraping and other coding efforts can be found in a repository on github: https://github.com/gushayki/SA-methods. For the sake of reproducibility, the last scraping was conducted on the 4th of April 2022. The corpus itself represents both a significant amount of work in collecting and cleaning it, and also a potentially useful resource for future work.

This full corpus was later split up into the following sets:

1. Training set - 90 reviews
2. Validation set - 500 reviews
3. Test set - 10,000 reviews
4. Reserve set - 18,494 reviews

The reasoning for these numbers is explained when elaborating on the process of creating the rule-based system later in this chapter. Different splits were utilised for the BERT neural network, and these are covered in Section 3.3.2.

All set splits aside from the Training set were done via the Pandas random sampling function [29]. The Training set was selected via a process detailed below in Section 3.2.1. The mean rating across the full corpus dataset is 7.6, and both the median and mode are 8.0. The full corpus is also heavily weighted towards positive reviews, though this is simply the result of the websites scraped rather than an intentional design: across the full dataset, if reviews are divided into *positive* (a rating of 5.1/10 or more) and *negative* (a rating of 5/10 or less), then 27,080 (93.1%) are *positive*, and 2004 (6.9%) are *negative*.

The randomly selected test set also has a mean rating of 7.6, median and mode of 8.0, and 93% positive vs. 7% negative reviews, and so has an almost identical statistical distribution of ratings.

## 3.2 Rule-based system construction

Detailed below is the full process for the construction of the rule-based system.

### 3.2.1 Searching for patterns

The first step was to examine examples from the corpus in detail to try to find potential patterns of language on which to base the resulting rule-based system. For this, 30 reviews were selected at random for three of the websites utilised (simply as it was felt that this was sufficient size for this task), Metal Crypt, Metal Underground and Metal Storm, giving a total of 90 reviews for this training subset. These groups of 30 were in turn subdivided into sets of 10: each consisted of 10 *negative* reviews (an overall reviewer-assigned score of 4/10 or lower), 10 *positive* reviews (an overall score of 7/10 or higher) and 10 *neutral* reviews (an overall score between 4.1 and 6.9). In this way, this subset was intended to be a representative spectrum of the sites used and the range of overall sentiment.

This set of 90 reviews is hereafter referred to as the *training set* (although it was not technically used for training any machine learning models, rather for the initial construction and then incremental refinement of the rule-based system).

Based on this manual examination, a number of avenues within this domain were identified that held potential for further work:

- **Aspects**: A very common style within this domain is to describe elements of the album, often with adjectives or other descriptive terms. A reviewer might talk about *the galloping guitars* or *the bombastic vocals*, etc. Analysis of these aspects specifically could prove an interesting method for assessing overall sentiment opinion.

- **Text flow**: The reviews, especially longer ones, often follow some form of pattern and flow. This is not consistent across all reviews by any means, but paragraphs within the text will often have some type of theme of primary point to make. For example, one paragraph may outline positive points, then it may be followed by one detailing the negative aspects. Alternatively, different paragraphs may cover different aspects of the music, or may proceed through the album in a more chronological style, covering each track one after another. This is in keeping with some of the work covered in Benamara, Taboada & Mathieu's *Evaluative Language Beyond Bags of Words* survey [21]. These sort of themes could help guide analysis of the reviews if they could be reliably detected and interpreted. One particularly common style is for the opening paragraph of a review to set the scene so to speak, outlining the band's history, etc. An argument could be made to bypass such introductory passages. This is covered in more detail in Section 3.2.4.3 below.

- **Hypotheticals**: Hypothetical statements, rhetorical questions, irony and sarcasm are all forms of writing that appeared within these reviews, and would likely present problems to any sentiment analysis attempt. However, this is a known challenge and field of NLP research all of its own, and to tackle it in this project would likely be too ambitious.

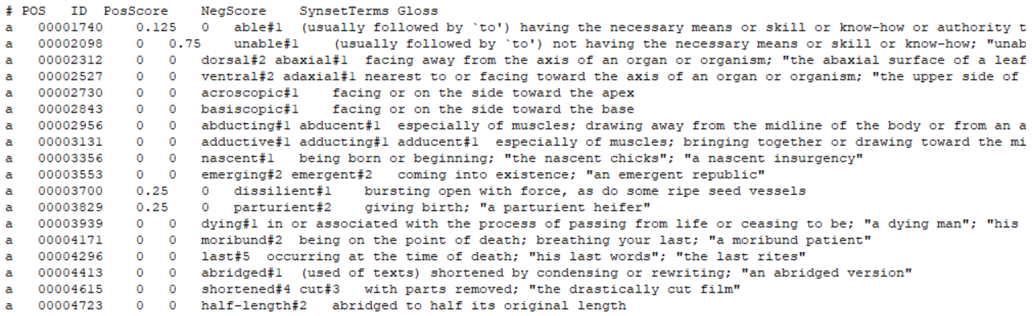
- **Comparisons**: A common technique of conveying sentiment within the reviews of this domain is to draw a comparison to a existing piece of work, providing an easy frame of reference for an experienced listener. A reviewer might, for example, compare the sound of an album to a well known band like Iron Maiden, or might describe an album's style as *[band A] meets [band B]*. However, while this is a common technique, it is used as often to simply describe the style or genre of a band as much as it is to convey sentiment. As such, it may not be the most reliable indicator.

- **Phrases**: Certain specific phrases occurred relatively frequently within the corpus sample, and these could be indicative of sentiment. For example, if a sentence refers to *the main problem* or *the biggest issue*, this alone indicates that there is such a problem/issue, without the need to fully understand what follows (though this could provide further information on it). Another example might be a reviewer asserting that the album is *their best since...* - the very fact that such a thing is noted conveys positive sentiment, regardless of how it is actually ranked by the reviewer. These phrases could be useful to note, though may also be context specific.

While each of the above could prove an interesting avenue for investigation, I decided to begin by looking into the *aspects* mentioned within reviews and how they might be used as indicators of sentiment, as this seemed to be a common linguistic device within these reviews, while also not being too overly ambitious to investigate and draw at least some conclusions on within the scope of this thesis project.

### 3.2.2 SentiWordNet 3.0.666

As outlined in Section 2.3, SentiWordNet (SWN) is a sizeable lexical database resource, assigning sentiment values to a huge range of words and terms. Its contents also go beyond simply marking words with a binary positive/negative classification, but assign values to represent the strength of the sentiment the word expresses.



```
# POS    ID  PosScore    NegScore    SynsetTerms Gloss
a    00001740    0.125   0   able#1  (usually followed by `to') having the necessary means or skill or know-how or authority t
a    00002098    0   0.75    unable#1    (usually followed by `to') not having the necessary means or skill or know-how; "unab
a    00002312    0   0   dorsal#2 abaxial#1  facing away from the axis of an organ or organism; "the abaxial surface of a leaf
a    00002527    0   0   ventral#2 adaxial#1 nearest to or facing toward the axis of an organ or organism; "the upper side of
a    00002730    0   0   acroscopic#1    facing or on the side toward the apex
a    00002843    0   0   basiscopic#1    facing or on the side toward the base
a    00002956    0   0   abducting#1 abducent#1  especially of muscles; drawing away from the midline of the body or from an a
a    00003131    0   0   adductive#1 adducting#1 adducent#1  especially of muscles; bringing together or drawing toward the mi
a    00003356    0   0   nascent#1   being born or beginning; "the nascent chicks"; "a nascent insurgency"
a    00003553    0   0   emerging#2 emergent#2   coming into existence; "an emergent republic"
a    00003700    0.25    0   dissilient#1    bursting open with force, as do some ripe seed vessels
a    00003829    0.25    0   parturient#2    giving birth; "a parturient heifer"
a    00003939    0   0   dying#1 in or associated with the process of passing from life or ceasing to be; "a dying man"; "his
a    00004171    0   0   moribund#2  being on the point of death; breathing your last; "a moribund patient"
a    00004296    0   0   last#5  occurring at the time of death; "his last words"; "the last rites"
a    00004413    0   0   abridged#1  (used of texts) shortened by condensing or rewriting; "an abridged version"
a    00004615    0   0   shortened#4 cut#3   with parts removed; "the drastically cut film"
a    00004723    0   0   half-length#2   abridged to half its original length
```

**Figure 3.3:** Examples of SentiWordNet entries. Each entry has a part-of-speech tag, an ID number, positive and negative scores, synset terms (those SWN entries considered synonymous) and a glossary briefly explaining the meaning.

Whatever precise method the rule-based system used, the substantial existing knowledge within SentiWordNet 3.0 would act as an excellent starting point for sentiment classification. However, as the project would be operating on a specific domain, modifications would likely be required for the existing SWN database, including the addition of new terms and the modification of existing entries. This can be for a variety of reasons:

- There are numerous terms utilised within this domain of hard rock/heavy metal reviews that SWN does not recognise by default, yet carry sentiment (for example, *headbanging*, which can be used as a descriptive term for effective rock or metal music).

- Some words may exist within the lexicon, but may have an additional meaning not encoded within SWN, or one particular meaning of the word may be more common within this domain than in general (for example, in this domain the word *heavy* is almost always going to refer to the musical style).

- Some words may exist within the lexicon and still be utilised in the same way, but may carry different sentiment within this domain than they would normally (for example, words such as *ominous* or *disturbing* are more likely to be used in a positive manner when describing heavy music).

To address the issue of entirely new words, a Python function was run across approximately 50% of the total corpus, returning unique words not recognised within SWN. This was done over approximately 50% of the corpus in order to save time, and because it was felt that if a word did not appear within that 50%, it was likely not sufficiently frequent to warrant a new entry within the lexicon. Given my personal familiarity with this genre of music, I considered myself sufficiently experienced to approximate the sentiment these new words would carry. Thus, I worked through the new words found within the corpus, noting all those which were likely to appear at least somewhat frequently, and whether they should be considered positive or negative.

In order to then come up with precise sentiment scores, I searched for the most similar term I could think of within existing SWN entries, and copied its scores across. For example, the words *trve*, *tr00*, *kvlt* and *cvlt* are all slang terms used in a derogatory fashion to refer to music (typically black metal) that is perceived as taking itself too seriously. As such, I decided to base its score on SWN entry #01688757, which refers to words such as *banal*, *trite*, *well-worn*, etc, "repeated too often; overfamiliar through overuse", and carries a positive score of 0.0 and a negative score of 0.375. This process, along with the creation of synonymous terms (*synset terms*), a short description and unique ID number, was repeated for all new terms, creating 90 new entries, which included both entirely new words and words with new, predominant definitions.

In the cases where the sentiment of an existing term needed to be changed, this was a simple process of amending the values within the SWN txt file. The list of words to amend in this fashion largely came from examination of reviews in the training set for such terms, and then synonymous words for them.

As this domain is one often characterised by aggressive and dark themes, this was most often a case of shifting negative values into positive ones. For these instances, positive and negative values were shifted to better represent the sentiment meaning of these words within this domain. For example, SWN 3.0 assigns negativity to the word *blasphemous*, but within this domain, this will usually be used in a more positive sense, given the commonplace use of antitheistic themes within the genre.

At other times, words were discovered with negative or positive sentiment in SWN which I felt should instead be neutral, and in these cases both the positive and negative scores were reduced to 0. For example, SWN 3.0 assigns negativity to various entries for the word *death*, but within the context of heavy metal music, this will most often describe lyrical themes, band, album or song titles, or musical style (*death metal*), and so a flat neutral value is better suited in this context. Similarly, the term *Gothic* will most often be used to describe the style of music, carrying no inherent positivity or negativity.

The modification of SWN also led on to the topic of word sense disambiguation: many words in SWN have more than one entry, representing different meanings. In order to modify the correct entry within SWN, I would need to know which entry needed to be modified in the first place. For example, the word *darkness* is one I wished to modify for this new, domain-specific version of SWN, as the term is often used in a non-negative manner to describe the atmosphere or lyrical content of music. But *darkness* has a total of 6 different synset entries in SWN. Some of these entries may be appropriate to modify for this domain (e.g. *darkness* meaning "absence of moral or spiritual values"), but others may be better left alone (e.g. *darkness* meaning " a swarthy complexion").

Which entry to modify depends on which version is selected by the disambiguation function (in this case, the pre-trained NLTK word sense disambiguation method). If this function is given a number of sentences typical of this domain containing the word *darkness*, and it is asked to disambiguate that word, which synset entry will it select? Will it always choose one consistently, or will it vary depending upon the sentence?

The NLTK WSD process was run over a series of artificial sentences created purely for this purpose, and the results were examined. An example of these artificial sentences, intended to investigate the word *malevolent*, is:

> *This album is malevolent. These songs are malevolent. This is malevolent music. Malevolent is how I would describe this. Every song on here is wonderfully malevolent.*

If the NLTK method reliably chose a single SWN entry for this word, that entry would be modified. Twelve words yielded less consistent results (*wickedness*, *heaviness*, *crush*, *smash*, *grind*, *anger*, *harshness*, *sadness*, *cacophony*, *roaring*, *violence* and *blasphemous*). In these cases, the different returned SWN definitions were examined closely and individual judgements were made. In some cases, multiple SWN entries would be modified, in other cases a single entry would be modified, as others did not seem likely to be the intended meaning for this domain (the word grind may often be used with this domain to refer to a particular musical quality, and the WSD returns definitions for both *the act of grinding to a powder or dust* and *press or grind with a crushing noise*, of which the second is the best fit, so this is the one modified). In some cases, it was decided that none of the existing entries in SWN fit, and instead a new entry should be made, as was done above for entirely new words (the word *heaviness*, for example, within this domain, will almost always refer to the *heavy* quality of the music, but such a definition does not exist in SWN 3.0).

There were also some cases where the word sense disambiguation method made a consistent choice, but it was clearly not a correct one. For example, with the word *fury*, the system consistently chose the synset entry *fury.n.04*, which refers to the Furies of Greek mythology. This seemed to be an issue with the disambiguation method rather than my own work, and only occurred in a few instances, so it was noted but left as is.

In total, 178 synsets were modified in this way within SWN. These modifications, along with the new additions, led to the creation of new, modified version of SWN for this project specifically. This was recorded as SentiWordNet 3.0.666, in keeping with the domain it is customised for. Details can be seen in Appendix A.1.

It should be noted that while the txt file used to store this information for the SWN lexicon could be modified, this was only effective when it came to the altered existing entries. Entirely new word entries could be created within the txt file, but would not be recognised when running SentiWordNet in Python with this new version of the file. For this, a manual *intercept layer* was created within Python: when a word is looked up, the system first checks a separate Python dictionary structure containing the new terms as keys and their respective positive/negative polarity as values. If the word is found in there, then the associated value is used as if it were looked up in SWN. If a word is not found within this dictionary of new terms, only then will the system check in the modified version of SWN.

### 3.2.3 Initial system construction

Based upon the observations made of the training set (see Section 3.2.1), it was decided to create a system that would seek out mentions of aspects of the albums within reviews, then assign sentiment to these aspects based on word dependency relations. These could be used to examine which elements of a review received the most praise or criticism, and in turn to output the sentences that impacted the most on sentiment scores to act as rationales for the system's ultimate predicted rating decision.

Reviews are split into sentences with NLTK, then each sentence is passed through the Stanza NLP pipeline, assigning part-of-speech tags and building word dependency trees to link words together. The rule-based system then searches for mentions of aspects based on a list of related terms intended to mark out when a reference to a particular element of an album is made:

- **Guitars**: ['guitar', 'guitarist', 'riff', 'riffing', 'lead', 'solo', 'soloing', 'noodling', 'riffage']

- **Vocals**: [’vocalist’, ’singer’, ’vocal’, ’vocalisation’, ’vocalising’, ’vokill’, ’singing’, ’scream’, ’screaming’, ’growl’, ’growling’, ’roar’, ’roaring’, ’shriek’, ’shrieking’, ’wail’, ’wailing’, ’snarl’, ’snarling’, ’bark’, ’barking’, ’howl’, ’howling’, ’yell’, ’yelling’, ’chant’, ’chanting’, ’croon’, ’crooning’, ’voice’, ’vox’]

- **Drums**: [’drummer’, ’percussionist’, ’drum’, ’drumming’, ’fill’, ’beat’, ’percussion’]

- **Production**: [’production’, ’mastering’, ’master’, ’mix’, ’production job’, ’producer’, ’recording’]

- **Breakdowns**: [’breakdown’]

- **Lyrics**: [’lyrics’, ’theme’]

- **Ambience**: [’ambience’, ’mood’, ’atmosphere’]

- **Writing**: [’composition’, ’structure’, ’writing’, ’songwriting’, ’arrangement’]

- **Bass**: [’bass’, ’bassline’, ’bassist’]

- **Melodies**: [’melody’, ’harmony’, ’rhythm’, ’grooves’]

- **Technicality**: [’technicality’, ’complexity’, ’signature’, ’skill’, ’talent’, ’musicianship’]

- **Keyboards**: [’keyboard’, ’synth’, ’piano’]

- **Symphonics**: [’orchestra’, ’choir’, ’symphonic’, ’orchestration’]

- **Exotic instruments**: [’sax’, ’saxophone’, ’tambourine’, ’trumpet’, ’flute’]

- **Creativity**: [’creativity’, ’experimentation’, ’variation’, ’variety’, ’diversity’, ’inspiration’]

- **Comedy**: [’humor’, ’humour’, ’comedy’]

- **Catchiness**: [’memorability’, ’catchiness’]

When aspect terms are found, the system then checks through the word dependency tree for all words linked to it, which includes both its *head* (the one above it in the tree) and any *dependent words* (those below it in the tree). These words are then checked against the modified version of SWN (and its *intercept layer*, see Section 3.2.3), and if these linked words have sentiment scores within, that sentiment is assigned to the appropriate aspect.

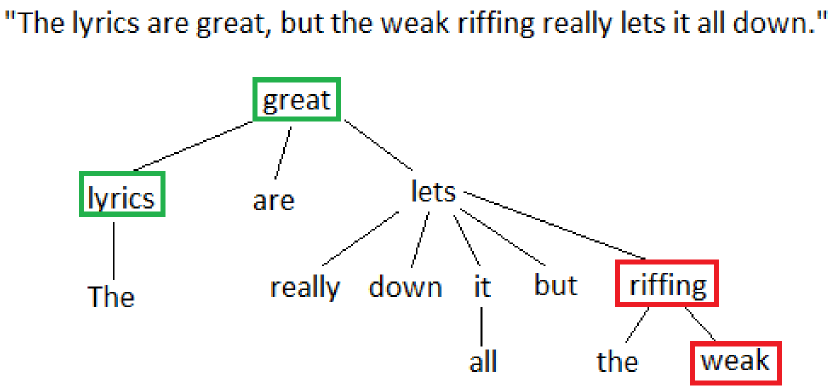

"The lyrics are great, but the weak riffing really lets it all down."
great
lyrics
are
lets
The
really
down
it
but
riffing
all
the
weak

**Figure 3.4:** Illustration of the rule-based system's basic core process (note: punctuation ignored in tree): the sentence is parsed to create the word dependency tree, and the words *lyrics* and *riffing* are identified as aspect terms. All words linked to each are checked. For *lyrics*, the words *The* and *great* will be checked in SWN, and the positive score of *great* will be recorded for the *vocals* aspect. For *riffing*, the words *lets*, the and *weak* will be checked in SWN, and the negative score of weak will be recorded for the *guitars* aspect.

An alternate version of the system was tested. Instead of searching for aspect terms and then connected sentiment terms, it would search for all sentiment terms and then connected aspect terms. However, this system achieved very little difference in its actual outputs while taking far longer to run, and so it was discarded early in the development process.

Two further modifications were made to the system, to account for hyphenated and multi-word terms and compounds.

Some of the sentiment terms added into this system consisted of multiple words separated by spaces, or hyphenated terms, and these would not necessarily be picked up on, as words linked by the parsing process may be separated. In order to combat this, each sentence is examined for these specific phrases, and if they are found, the words that make them up are merged into one single word (e.g. *by-the-numbers* is converted into *bythenumbers*), which the *intercept layer* above SWN is encoded to look for. These terms appear to still be handled correctly by the Stanza pipeline, and so this modification is not believed to have any negative impact on the process, while allowing these terms to be more accurately picked up on.

The Stanza pipeline is also capable of identifying compound terms within a sentence. For example, the terms *production job* or *guitar work* refer to the production and guitar aspects, respectively, but sentiment terms may be linked to *job* or *work*, and thus would be missed by simply searching for words linked only to *production* or *guitar*. When Stanza identifies these adjective-noun compounds, it marks one of the words with *compound* in the *deprel* field such that the *head* of the compound word links the aspect and sentiment terms together. As such, when the system checks for aspect terms, it also checks for this marker. If it is found, then the search for linked words will be extended to include those linked to the *head* of the compound word as well, as illustrated in Figure 3.5.

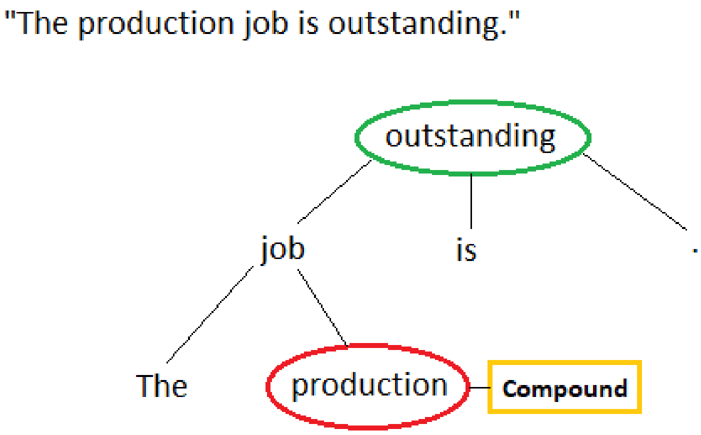


**Figure 3.5:** An illustration of compound links within Stanza parsed text. In the above, the sentiment term *outstanding* is removed from the aspect term *production*, and so would not normally be picked up on. *Production job* is recognised as a compound term however, and so *production* is given a *compound* marker. When the system detects this, it extends its search for linked sentiment terms to search not just for those words linked to *production*, but any words linked to the head of *production*, in this case *job*, allowing it to find *outstanding*.

### 3.2.4 Further additions & refinement

Expansions and additional components were added to the system over time, with the system being run repeatedly across the training set as it was modified. This revealed some more minor ways in which the system could be incrementally improved upon (further modifications to make to the new version of SWN, for example). But there were also a number of more substantial additions made to the system, which will be detailed below. As with the construction of the system above, all observations and testing done to develop the system was done with the training set of 90 reviews.

#### 3.2.4.1 Contextual polarity

Contextual polarity refers to polarity that is imparted not from a word alone, but from the context that word is in, i.e. what lies around it. This can include negation terms (e.g. *not*, *never*) as well as words that boost or dampen the impact of other sentiment terms (referred to by Taboada et al. as *intensifiers* and *downtuners* [14]), such as *very* or *slightly*. Implementing this at some level within the rule-based system was expected to lead to improved results, as saying an aspect of an album is *very good* should naturally count for more than saying it is only *slightly good*, while saying it is *not good* should actually be considered negatively.

For this, inspiration was taken from the handling of these terms within SO-CAL (see Section 2.3.4). A list of modifier terms and associated float values was taken from SO-CAL and encoded into the system as a dictionary structure [30], with the modifier term being the key and the float value representing the percentage increase or decrease that the modifier would apply to a sentiment term linked via the parsing. Some negation terms were also added to this in the same fashion.

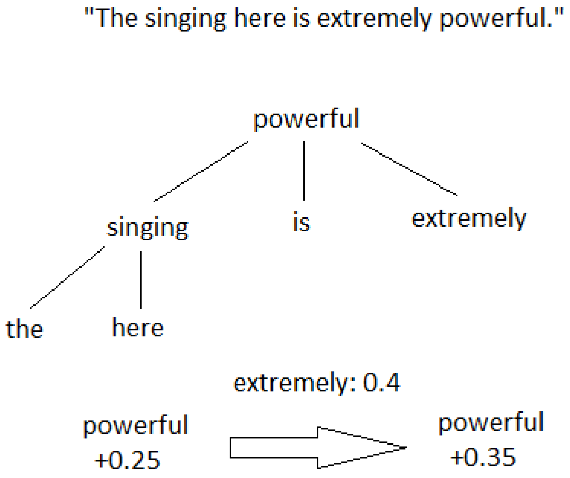


**Figure 3.6:** In the above example (note: punctuation removed from tree), the sentiment term *powerful* is applied to the *singing* aspect (*vocals*). The system checks for words linked to the sentiment term for modifiers. In this instance, it finds *extremely*, which increases the value of *powerful* by 40%. *Powerful*'s positive score of 0.25 is increased to 0.35.

The use of an n-grams approach (as detailed in Section 2.1.2) to search for linked modifier terms was tested, but this was not found to produce significantly better results than searching through the word dependency trees, and was felt to be counter to this system's theme of utilising the parsed results.

#### 3.2.4.2 Implicit aspects & miscellaneous sentiment

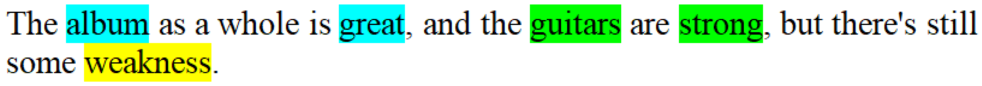
Examination of the results of the rule-based system in its early form revealed that in many cases, sentiment expressions were being missed because they were not directly linked to a recognised aspect term. In many cases, these expressions were referring to the album as a whole, which overlaps with the notion of implicit aspect/topic evaluation [21]. This is where sentiment is expressed, but the target is not stated directly and has to be inferred (e.g. "It is really good" - In order to properly interpret this, we need to know what *it* is referring to here). While not every instance of referring to the album as a whole was strictly implicit evaluation in this sense (sometimes *the album* would be stated clearly), for the purposes of this system this is referred to as the implicit aspect, in contrast to the explicit aspects detailed above. Thus, we want to capture sentiment expressions referring to songs or the album as a whole, and logically these should be given greater weight than any one explicit aspect alone, as they refer to the entire product.

Accounting for this implicit aspect was treated in the same way as the explicit ones, where a list of synonyms was added for what would be typically taken as a reference to the album as a whole, and linked words were checked against the modified version of SWN. The synonyms for the implicit aspect were:

- ’song’
- ’album’
- ’track’
- ’sound’
- ’band’
- ’artist’
- ’group’
- ’musician’
- ’performer’
- ’member’
- ’music’

In the case of the implicit aspect, if the above synonyms are found, they are also checked to ensure they are marked as nouns by the Stanza pipeline part-of-speech tagging. This is to ensure that these more general words are indeed being used to refer to the album/songs as a whole, rather than in some other sense (e.g. if a reference is made to "the group", then this is likely to refer to the band in question, but if it is used as a verb "to group together", this is likely not the case).

Even when these are captured, there also remain some other sentiment expressions not attached to any aspects, implicit or explicit, by the parsing. These represent valuable data that we do not wish to discard entirely, and so the system was expanded to make an additional pass through each sentence after checking for implicit and explicit sentiment. The index values of the words linked to aspects would be recorded on the first pass, then all remaining words (aside from the aspect terms themselves) would be checked against SWN. Any sentiment found in this way would be recorded as an additional *miscellaneous* or *unattached sentiment* score. An example of this can be seen in Figure 3.7.



The album as a whole is great, and the guitars are strong, but there's still some weakness.

**Figure 3.7:** In the above, *great* will be linked to *album* for the implicit aspect, and *strong* will be linked to *guitars* for an explicit aspect, but *weakness* is tied to neither, and so will be picked up in the additional pass for unattached sentiment.

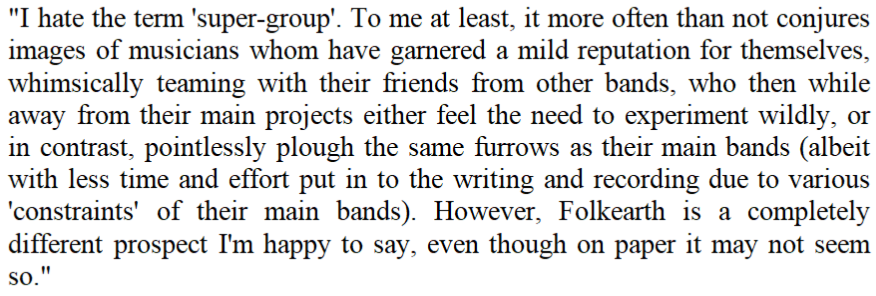
Both the implicit aspect and unattached sentiment scores could be factored into the final calculations used by the system to make rating predictions. Also similarly to the explicit aspects, limits could be placed onto these, though greater values were used than for each individual explicit aspect, as it was likely that there would be a greater number of sentiment expressions about the album as a whole, or unattached sentiment, than any single explicit aspect.

#### 3.2.4.3 Ignoring the introduction

As mentioned in Section 2.2, discourse analysis is an avenue of potential future interest for NLP. This refers to treating a sentiment document as a form of discourse between the writer and their audience, and recognising a certain flow within the document as a result, which can include sentiment-relevant patterns.

While assessing the Training set reviews manually, a certain degree of this sort of flow was noticeable, though it came in varied forms, if at all. In some cases a paragraph may focus on a particular aspect of an album (e.g. the production). At other times paragraphs could reasonably be said to be generally positive or negative, with the reviewer alternating between these across the review. In some instances, the review took a more chronological approach, going through the album song by song.

While some patterns of these sorts could be discerned, in general they did not seem consistent enough to reliably build functionality into the system based on them, at least not within a project of this timescale. However, one element of flow within the reviews did seem consistent enough to warrant operating on its assumed presence: that of an introductory section to the review. While not present in every single instance, reviews would frequently dedicate an opening segment to introducing the band and describing their history and past releases. Sentiment expressed in such passages would often not relate directly to the album being reviewed currently, as can be seen in some examples in Figure 3.8.



"I hate the term 'super-group'. To me at least, it more often than not conjures images of musicians whom have garnered a mild reputation for themselves, whimsically teaming with their friends from other bands, who then while away from their main projects either feel the need to experiment wildly, or in contrast, pointlessly plough the same furrows as their main bands (albeit with less time and effort put in to the writing and recording due to various 'constraints' of their main bands). However, Folkearth is a completely different prospect I'm happy to say, even though on paper it may not seem so."

**Figure 3.8:** The introductory paragraph from the review *Folkearth - A Nordic Poem*, Review 9240. As can be seen, there are many potential sentiment expressions, but the reviewer is outlining a general position of theirs rather than referring to the album in question. Only in the last sentence is there a statement made about this album specifically, and this sentiment is expanded upon later in the review.

On the occasions when sentiment regarding the current album was expressed within introductory sections, it would frequently be repeated later on in the review, serving a function not dissimilar to an abstract at the start of an academic paper: briefly summing up a major point that would be expanded upon further in.

This gave rise to the thought that such introductory sections could, perhaps even should, be ignored when assessing sentiment within this domain. If an introduction talked about the band's previous album being good or bad, this would not have any bearing on the current album. As such, this was encoded into the rule-based system. An *intro_to_ignore* variable was added, which would specify the percentage of the total number of sentences in a review to be considered the *introduction*, and consequently removed from the review before assessing it for sentiment. The number resulting from the calculation would always be rounded downwards to ensure that not too much of the review was removed. The use of this method, utilising a percentage and rounding down, also ensured that shorter reviews, which may not have a dedicated introductory section, could still be assessed in full. With an intro_to_ignore value of 5%, for example, only a review of 20 sentences or more would have anything removed.

### 3.2.5 Final version

After the above additions were made, the system's full process was as follows:

1. The review is split into sentences using the NLTK tokeniser, simply to allow the following processing steps to take place sentence-by-sentence. Each review is also assigned a dictionary listing all the possible aspects and an overall sentiment score for each (each starting at 0). An intro_to_ignore value ($i$) is also set. If $i > 0$, the first $i$% of sentences in the review will be removed (always rounding down). At this stage, sentences will also be checked to compress certain hyphenated or multi-word sentiment terms and modifier terms,

as outlined above in Section 3.2.2.

2. Part-of-speech tagging, lemmatisation and parsing is then conducted on the sentence using the Stanza NLP pipeline. This splits and lemmatises the words and builds dependencies between them, allowing for the further steps.

3. Each sentence will be checked for all aspects and their synonyms, both implicit and explicit. If none are found, the system skips ahead to step 5 for this sentence.

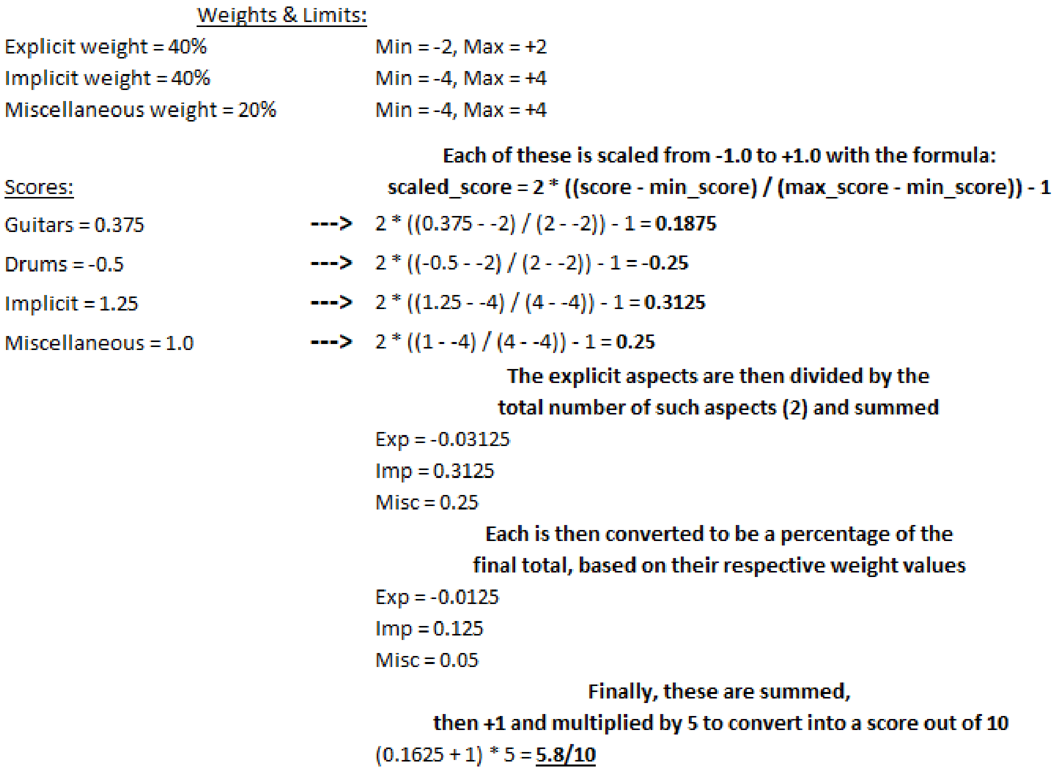
4. If an aspect is found, the word dependencies are searched for linked words; that is, words which have the aspect as their *head*, and the word that is this aspect's *head* (and possibly more if compounds are found, as detailed above). If these linked words are found within the modified version of SWN (or the intercept layer of new terms), then the appropriate sentiment scores are assigned to that aspect's sentiment score for this review. Linked contextual polarity terms will also be factored in to calculations of sentiment scores. Word Sense Disambiguation is also conducted to ensure the correct version of a word (and its assigned SWN score) is used. If multiple aspects are found within a single sentence, this process will be repeated, moving on to the next aspect each time.

5. Once all explicit aspects, and the implicit aspect, have been checked as above, the sentence will be scanned once more for any remaining sentiment terms, excluding the indices of words already checked in connect with implicit/explicit aspects, and the aspect terms themselves. This unattached sentiment is recorded as its own on-going score. The sentence itself will also be assigned a score for all sentiment found within through steps 4 and 5.

6. Once all sentences have been processed in this manner, we are left with sentiment scores for the different possible aspects (including implicit, and unattached sentiment), along with scores for each sentence. The scores of the implicit aspect, each individual explicit aspect, and the unattached sentiment across the review, are each limited, but cut to these limits if they exceed them over the course of the review. This is both due to the intuitive notion that repeated praise/criticism of a single element should not keep on counting for more sentiment score, and also for the practical reason that upper and lower limits are needed to scale the scores and make an overall prediction. The sentences with the highest positive or negative impact on each aspect (i.e. the greatest departure from the neutral point of 0) can be produced as rationales justifying the final predicted rating. The explicit aspect scores can also be taken as indicators of what are considered the best and worst aspects of an album by this review.

7. Lastly, the system makes a prediction of an overall sentiment score for the full review. The implicit aspect score, unattached sentiment score and each explicit aspect score will each be scaled from -1 to +1. Thus, the starting point

of 0 for each aspect is considered neutral, while -1 and +1 would indicate the maximum amount of negativity or positivity for that respective aspect. The explicit aspect scores will be summed up and divided by the number of aspects that were scored (i.e. did not remain at 0). The three scores (explicit, implicit, unattached) will then be modified based on how much weight each is assigned, and summed up to produce a final score on the -1 to +1 scale. By adding 1 and multiplying by 5, we then convert this to a predicted rating out of 10.



Weights & Limits:

Explicit weight = 40% Min = -2, Max = +2

Implicit weight = 40% Min = -4, Max = +4

Miscellaneous weight = 20% Min = -4, Max = +4

**Each of these is scaled from -1.0 to +1.0 with the formula:**

**scaled_score = 2 * ((score - min_score) / (max_score - min_score)) - 1**

Scores:

Guitars = 0.375 ---> 2 * ((0.375 - -2) / (2 - -2)) - 1 = **0.1875**

Drums = -0.5 ---> 2 * ((-0.5 - -2) / (2 - -2)) - 1 = **-0.25**

Implicit = 1.25 ---> 2 * ((1.25 - -4) / (4 - -4)) - 1 = **0.3125**

Miscellaneous = 1.0 ---> 2 * ((1 - -4) / (4 - -4)) - 1 = **0.25**

**The explicit aspects are then divided by the total number of such aspects (2) and summed**

Exp = -0.03125
Imp = 0.3125
Misc = 0.25

**Each is then converted to be a percentage of the final total, based on their respective weight values**

Exp = -0.0125
Imp = 0.125
Misc = 0.05

**Finally, these are summed, then +1 and multiplied by 5 to convert into a score out of 10**

(0.1625 + 1) * 5 = **5.8/10**

**Figure 3.9:** Illustration of calculations done within the rule-based system to come to a final rating prediction. Each type of aspect (explicit, implicit and unattached sentiment, labelled here as miscellaneous) has a weight and a set of limits. In the above example, the explicit aspects guitars and drums have received scores of 0.375 and -0.5 respectively, while the total implicit score is 1.25 and the unattached sentiment is worth 1.0. These are scaled to -1 to +1, the explicit aspects are divided by their total number and summed, each value is modified by its weight, then they are all summed, have 1 added and are multiplied by 5 to convert to an out-of-10 rating.

## 3.3 Other systems

In addition to the rule-based system, two other systems were developed for comparative evaluation: a simple sentiment counter and a neural network. Substantially less time was spent on development of these systems, as they were only ever intended as points of comparison for evaluation. The simplistic system was (as its name suggests) intended to operate in a very straightforward manner, and for the neural network, the pre-trained BERT model was utilised rather than something constructed from scratch.

### 3.3.1 Simplistic sentiment counter system

This system was constructed specifically to represent taking a much simpler approach to sentiment analysis, and to assess whether the additional time spent in building and refining a lexicon of rules actually results in better performance than something more basic. It attempts to answer the question of whether just looking for sentiment terms within a review is sufficient to make a reasonable prediction of overall sentiment across the document, even if we ignore many linguistic nuances and details.

In this system, sentences are still tokenised and passed through the Stanza pipeline, and these words will be disambiguated and looked up in SWN (utilising version 3.0.666, as the rule-based system does) one by one to see what sentiment scores they have. However, no attention is paid to word dependencies, aspects or contextual polarity modifiers, nor is any of the introduction of a review ignored. Instead, this system takes only the most simple of approaches: it looks up every word in the review and keeps a count of the overall sentiment scores found throughout the review, then converts this into a rating prediction in the same manner as the rule-based system does (though with no issue of weights for different aspect types). This count of total sentiment still utilises a limit, but this is fixed at -10 to +10, representing the sum of the limits on the three types of aspect assessed within the rule-based system.

### 3.3.2 BERT

For the implementation of BERT as a neural network, a pre-existing Python script was utilised, supplied by the thesis supervisor. This script was able to make use of a pre-trained BERT model from the Hugging Face library. In this case, bert-base-cased was used, an English language model with case sensitivity [34].

The pre-trained base model is loaded, then fine-tuned with a designated training set and validation set (see below for details of different sets on several separate runs). This fine-tuning training is done as a supervised regression task: the model (including its regression output layer) is shown review texts as inputs, learning contextual meanings relevant to this domain and associating them with the corresponding ground truth label (the actual rating of each review). It thus learns to make associations between the texts and the ratings given. Once fitted in this fashion, the model can be given a new, unlabelled review text, will assess the text based on its learning, and output a number ranging from 0.0 to 10.0 as a regression prediction of its rating.

In total, three separate runs were made with this model, each time fine-tuning a fresh instance so that previous learning would not be carried forward:

1. First, the model was fine-tuned on only the same training set of 90 reviews that was used for development of the rule-based system, and validated with the corresponding validation set of 500 reviews. This was done to provide as close of a match as possible to what the rule-based system was constructed with. It was then tested on the matching test set of 10,000 reviews.

2. In the second run, BERT was fine-tuned and validated using the reserve set of 18,494 reviews, split into a training set of 14,800 and a validation set of 3694 (an approximate 80/20 split of the reserves). This split was done with the random sampling method [29]. It was then tested on the same test set of 10,000 reviews.

3. Finally, the test set was altered using a process of selective masking. The full test set was first passed through the finished rule-based system (which had already been through a process of hyperparameter fine-tuning, detailed in Section 3.4 below), and the system output indices for the three sentences it deemed to have the greatest sentiment weight in each review. These sentences were then removed (masked) from the test set, and it was then passed through BERT as in the second run. This was done to see what impact this would have on BERT's predictions, which could provide some insight into how it judges sentiment in comparison to the known process of the rule-based system.

It should be noted that in some 30 instances, the full review was fewer than four sentences long, and so the selective masking would remove the entire review. In this cases, the least weighty of the three masked sentences was added back in manually, so in effect only the two weightiest sentences were masked for these reviews.

## 3.4 Hyperparameter tuning

Once the systems had reached their finished forms, there still remained the issue of their hyperparameters. These are elements of a system that can have their values modified, and may affect the performance of the system depending on what the values are set as. Such values should ideally be fine-tuned in order to find the set that produces the optimal performance.

Within the rule-based system there were a number of such elements. The quantity of a review to be removed as an introduction section (if any) could vary. The three categories of sentiment aspect (the implicit aspect, the explicit aspects combined, and the miscellaneous, unattached sentiment) could also be varied in how they were balanced: they would each contribute to the final predicted rating in some capacity, but more weight could be assigned to any one of them at the cost of weight from the others. In mathematical terms, if their respective weights were represented as percentages, they would all need to sum to 100, but could be balanced in any proportion within that. Since there was this constraint of a maximum sum of 100%, the weighting of these three elements was considered to be a single hyperparameter to be tuned, with each possible value representing a different weight distribution (e.g. an implicit/explicit/miscellaneous split of 40/30/30 would be one possible value, a split of 30/40/30 would be another, and so on).

The limits placed upon each of these same three types of aspect (implicit, explicit, unattached) could also be fine-tuned, but it was decided to leave these as fixed values (-2 to +2 for each explicit aspect, -4 to +4 for implicit, and the same for unattached). This was done primarily as it was felt that the effect of changing these values would ultimately be achieved through tuning the weight distributions as outlined above, just in a different way. Treating these limits as separate hyperparameters to tune would also have substantially increased the time needed to find the optimal set of hyperparameter values.

In order to find the optimal values for the introduction to ignore percentage and the sentiment aspect weight distribution, a function was coded in Python that would try different hyperparameter values across a given set of reviews, and would return the metric results so that the best could be selected. This acted as a custom grid search of sorts, though in order to save time, the two hyperparameters were treated naively (i.e. they were fine-tuned separately. In order to be entirely thorough, each possible combination of values for the two hyperparameters should be tested, as the 'best' value for one in isolation may not remain the best when the other hyperparameter changes. However, this dramatically increases the time taken to search through all possible combinations).

The following metrics are utilised:

- **MAPE** - Mean Absolute Percentage Error. This metric gives a percentage that represents the average difference between the predicted values and actual values. The smaller this percentage, the better the results. As a percentage metric, this is most useful when the dataset values are relatively narrow in range, as extreme values can lead to very high percentages.

- **RMSE** - Root Mean Square Error. This metric calculates the root of the average of the squared errors across the dataset, being the standard deviation of the residuals. The smaller this value, the better the results.

- **Spearman's rho correlation coefficient** - This metric measures the rank correlation between two variables, in this case between the predicted ratings and actual ground truth values. It can thus be taken to represent the correlation between these two. While it is very common to measure sentiment analysis tasks with accuracy, this makes less sense with a numerical scale, where a prediction of 9.0 is closer to 10.0 than a prediction of 2.0 would be. Accuracy measures fail to see this distinction, while correlation-based measures don't [40]. A value of 0 for this metric indicates no correlation whatsoever, while values of -1.0 or +1.0 indicate strongest correlation (inverted correlation in the case of -1.0).

- **P-value** - This metric, used in conjunction with the correlation test, represents the probability of obtaining results as extreme as those being observed within the test, assuming the null hypothesis is correct. The smaller this value, the less likely it is that these results would be observed under the null hypothesis.

For fine-tuning these hyperparameters, the validation set of 500 reviews was used. The assessment of the results was done using Spearman's rho correlation coefficient to measure correlation between the predictions made by the system and the actual ground truth values.

For the introduction-to-ignore hyperparameter, values ranging from 0% to 20% were tested, increasing in 2% increments. It was felt that more than 20% of a review could not reasonably be deemed an *introduction*, and increments of 1% would only make a difference with reviews of more than 50 sentences long. As detailed in Table 3.2, the best single performance came with ignoring the first 8% of sentences in a review. While this did not make a large difference overall, the consistency of the Spearman's rho values alone is interesting to note, as it suggests that even if ignoring more of the introduction of a review does not actively *improve* the results much, it also does not *harm* them either. This in turn seems to support the hypothesis that the introduction of these reviews is not especially important in establishing the sentiment of the review.

**Table 3.2:** Table of intro-to-ignore hyperparameter search results. P-value not reported as it was always lower than 0.001.

| intro_to_ignore | MAPE | RMSE | Spearman's rho |
|---|---|---|---|
| **0%** | 0.254 | 2.101 | 0.369 |
| **2%** | 0.254 | 2.101 | 0.369 |
| **4%** | 0.254 | 2.102 | 0.370 |
| **6%** | 0.253 | 2.105 | 0.372 |
| **8%** | 0.254 | 2.105 | **0.378** |
| **10%** | 0.256 | 2.116 | 0.377 |
| **12%** | 0.257 | 2.123 | 0.374 |
| **14%** | 0.258 | 2.131 | 0.370 |
| **16%** | 0.259 | 2.144 | 0.365 |
| **18%** | 0.260 | 2.150 | 0.366 |
| **20%** | 0.263 | 2.168 | 0.369 |

For the distribution of weights hyperparameter, a total of 36 variations were tested. This represented the full spectrum of possible values with 10% increments where each of the three component weights must weigh at least 10%. The full results can be seen in Table 3.3.

**Table 3.3:** Table of weight distribution results. P-value not reported as it was always lower than 0.001.

| **Weight distribution (implicit/explicit/misc)** | **MAPE** | **RMSE** | **Spearman's rho** |
|---|---|---|---|
| 10/10/80 | 0.244 | 2.130 | 0.318 |
| 20/10/70 | 0.223 | 1.979 | 0.334 |
| 30/10/60 | 0.210 | 1.873 | 0.342 |
| 40/10/50 | 0.205 | 1.819 | 0.353 |
| 50/10/40 | 0.209 | 1.822 | 0.365 |
| 60/10/30 | 0.220 | 1.875 | 0.368 |
| 70/10/20 | 0.235 | 1.985 | 0.366 |
| 80/10/10 | 0.254 | 2.139 | 0.349 |
| 10/20/70 | 0.222 | 1.976 | 0.321 |
| 10/30/60 | 0.208 | 1.871 | 0.329 |
| 10/40/50 | 0.205 | 1.823 | 0.336 |
| 10/50/40 | 0.219 | 1.839 | 0.341 |
| 10/60/30 | 0.226 | 1.917 | 0.343 |
| 10/70/20 | 0.246 | 2.051 | 0.339 |
| 10/80/10 | 0.268 | 2.223 | 0.320 |
| 20/70/10 | 0.263 | 2.173 | 0.352 |
| 30/60/10 | 0.258 | 2.135 | 0.370 |
| 40/50/10 | 0.254 | 2.105 | 0.378 |
| 50/40/10 | 0.252 | 2.094 | 0.376 |
| 60/30/10 | 0.250 | 2.093 | 0.370 |
| 70/20/10 | 0.251 | 2.106 | 0.364 |
| 20/20/60 | 0.208 | 1.864 | 0.342 |
| 30/20/50 | 0.202 | 1.803 | 0.356 |
| 40/20/40 | 0.205 | 1.801 | 0.367 |
| 50/20/30 | 0.217 | 1.853 | 0.376 |
| 60/20/20 | 0.232 | 1.957 | 0.374 |
| 20/30/50 | 0.202 | 1.805 | 0.351 |
| 20/40/40 | 0.207 | 1.809 | 0.359 |
| 20/50/30 | 0.221 | 1.878 | 0.362 |
| 20/60/20 | 0.240 | 1.999 | 0.368 |
| 30/50/20 | 0.235 | 1.967 | 0.382 |
| **40/40/20** | 0.233 | 1.951 | **0.385** |
| 50/30/20 | 0.232 | 1.947 | 0.381 |
| 30/30/40 | 0.205 | 1.796 | 0.367 |
| 30/40/30 | 0.217 | 1.854 | 0.375 |
| 40/30/30 | 0.216 | 1.846 | 0.378 |

As can be seen above, the best performance came with a distribution of 40/40/20 for implicit, explicit and miscellaneous. This, along with 8% introduction to be ignored, was taken as the optimal value to work with in future runs of the rule-based system.

The simplistic system did not have any hyperparameters to tune, and the BERT network was fine-tuned but otherwise used in its default set-up.

# 4

# Results

This chapter presents the results of running all systems through the full test set of the corpus, as well as results from other specific points of interest.

## 4.1 Primary metrics

In total, three systems were run with various different settings and sets:

1. The **rule-based system** was developed using the training set of 90 reviews and the hyperparameters were tuned using the validation set of 500 reviews.

2. The **simplistic system** was developing using the training set of 90 reviews. It required no specific hyperparameter tuning, so no validation set was needed.

3. The **BERT neural network** was run three times, with the model started fresh each time. In the first pass, BERT was fit using only the training set of 90 reviews and validated with the validation set of 500, in order to provide a more straightforward comparison point for the other two systems. In the second and third passes, BERT was fit using the reserves set, fit on a randomly sampled subset of 14,800, then validated with the remaining 3694 reviews (an approximate 80/20 split of the reserves).

Each of the above was tested using the same test set of 10,000 reviews. In the third run with the BERT NN, the reviews within this test set were selectively masked, as detailed in Section 3.3.2.

While each system was run for the full 10,000 test set, the evaluation metrics were based on 9985, as 15 reviews had an actual rating of 0.0, which would present a problem when trying to perform MAPE evaluation. As such, these 15 results were simply removed when checking the metrics.

In addition to the metrics explained in Section 3.4, both the actual rating and predicted rating are translated into a binary categorical division, *positive* (score $> 5.0$) or *negative* (score $< 5.1$). These can then be used to generate a confusion matrix showing the number of True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) results. These can then be used to calculated Precision, Recall and F-score results. In this case, these metrics are used instead of simple accuracy because the corpus and test set are imbalanced: within both, approximately 93% of all reviews fit the above definition of positive.

- **Precision** ( TP/(TP+FP) ) indicates how many of the sentiment categories predicted by the system are correct.

- **Recall** ( TP/(TP+FN) ) indicates how many of the actual sentiment categories the system is picking up on.

- **F-score** ( 2 * (precision*recall)/(precision + recall) ) aims to balance precision and recall.

**Table 4.1:** Confusion matrices for all 6 systems/setups. Note that as the negative sentiment category is the less common one, it is treated as the "positive" result, such that a True Positive will occur when both the prediction and ground truth values are negative sentiment. Thus, in each 2x2 grid above, the top left represents TP, the top right FN, the bottom left FP and the bottom right TN.

| **Rule-based (modified SWN)** | | |
|---|---|---|
| | **Predict neg** | **Predict pos** |
| **Actual neg** | 242 | 439 |
| **Actual pos** | 1003 | 8301 |

| **Rule-based (original SWN)** | | |
|---|---|---|
| | **Predict neg** | **Predict pos** |
| **Actual neg** | 316 | 365 |
| **Actual pos** | 2273 | 7031 |

| **Simplistic** | | |
|---|---|---|
| | **Predict neg** | **Predict pos** |
| **Actual neg** | 208 | 473 |
| **Actual pos** | 980 | 8324 |

| **BERT (small fit)** | | |
|---|---|---|
| | **Predict neg** | **Predict pos** |
| **Actual neg** | 681 | 0 |
| **Actual pos** | 9304 | 0 |

| **BERT (large fit)** | | |
|---|---|---|
| | **Predict neg** | **Predict pos** |
| **Actual neg** | 126 | 555 |
| **Actual pos** | 77 | 9227 |

| **BERT (selective masking)** | | |
|---|---|---|
| | **Predict neg** | **Predict pos** |
| **Actual neg** | 135 | 546 |
| **Actual pos** | 132 | 9172 |

**Table 4.2:** Metric results for all three systems. P-value not reported as it was always lower than 0.001.

| **System** | **MAPE** | **RMSE** | **Spearman's rho** | **Precision** | **Recall** | **F-score** |
|---|---|---|---|---|---|---|
| **Rule-based (modified SWN)** | 24.2% | 1.993 | 0.314 | 0.194 | 0.355 | 0.250 |
| **Rule-based (original SWN)** | 28.7% | 2.411 | 0.238 | 0.122 | 0.464 | 0.193 |
| **Simplistic** | 23.9% | 1.964 | 0.337 | 0.175 | 0.305 | 0.222 |
| **BERT (small fit)** | 38.8% | 3.114 | 0.072 | 0.068 | **1.0** | 0.127 |
| **BERT (large fit)** | **16.6%** | **1.239** | 0.497 | **0.621** | 0.185 | **0.285** |
| **BERT (selective masking)** | 16.7% | 1.254 | **0.499** | 0.506 | 0.198 | **0.285** |

## 4.2 Other points of interest

In addition to the comparison of the different methods above, some other connected avenues were also investigated as potential points of interest for the project. While these do not directly contribute to the metrics outlined in Section 4.1, they can provide some insights into related points.

### 4.2.1 Sentiment flow

As outlined in Section 2.2, discourse analysis is one potentially interesting avenue of future development for sentiment analysis [21]. This refers to the treatment of a sentiment document as a piece of discourse between writer and reader, and the assumption that there may be some sort of sentiment-relevant flow to the text as a result.

To investigate this in detail would be a substantial project of its own, and it was only implemented within the rule-based system here in a single, shallow manner, that of ignoring a designated *introduction* section of a review and not factoring it into the sentiment analysis. However, I also felt that it could be interesting to plot the flow of the sentiment across the sentences of a review, to see if any common patterns could be discerned across different reviews. Even though this would not impact on the results in this instance, it could nevertheless reveal some insight into the topic of discourse analysis and how it might be expressed within this particular domain.

For this, I decided it would be best to examine examples where the rule-based system performed best, as these would represent the most meaningful results. The 50 best performance results were selected, those being the instances where the difference between the rule-based system's prediction and the actual rating was 0.0. The sentiment scores of each sentence within each review was then plotted, to represent the flow of the sentiment within the review text.

Note that one of the 50 reviews was replaced (with another that also had a prediction-actual rating difference of 0.0) as it featured language I did not wish to include within this report and its associated documents. Similarly, one review among the 50 had a word censored as it may be triggering for some readers.

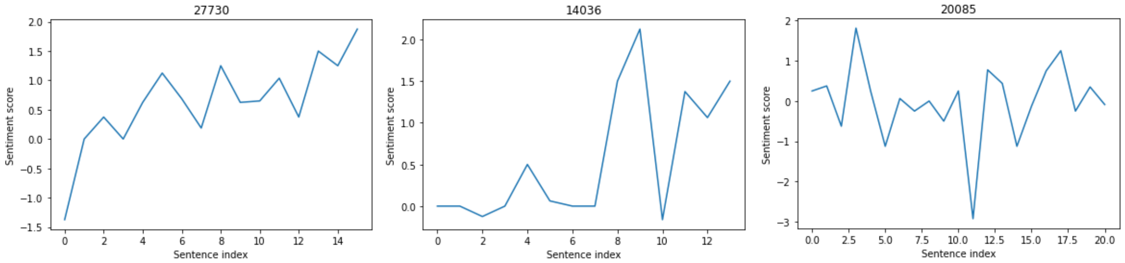


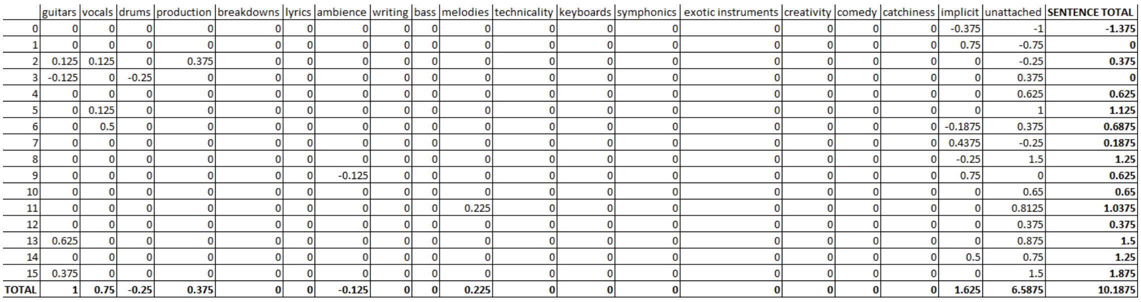
**Figure 4.1:** Some examples of sentence sentiment plots.

### 4.2.2 Explicit aspect polarity

While the rule-based system had various components and additions made over the course of its development, its focus was consistently on the notion of different aspects of the music and albums, and how reviewers would frequently frame their analysis of these albums around these aspects. The reviews that made up the domain included overall numerical ratings to represent the sentiment of the reviewer, and the systems above would each make predictions of these overall ratings. However, the rule-based system would also, as part of its process, score individual aspects throughout the review, and it was felt that these themselves could provide an interesting avenue for examination.

Unlike the overall ratings, these aspect ratings have no corresponding ground truth label among the corpus (some review sources do include a breakdown of their overall rating, such as on Metal Storm, which assigns scores for *performance*, *songwriting*, *originality* and *production*, but this is not consistent across sources, nor does it match up to this project's manner of dividing up aspects). As such, in order to assess these individual aspect ratings, manual annotation was required.

The same set of 50 high-performance results was used here as for the plotting of sentiment flow across reviews in Section 4.2.1, where the difference between prediction and ground truth was 0.0, and so the aspect scores within these reviews would be the most reflective of genuine reviewer sentiment. These results are examined in Section 5.2.2.



| | guitars | vocals | drums | production | breakdowns | lyrics | ambience | writing | bass | melodies | technicality | keyboards | symphonics | exotic instruments | creativity | comedy | catchiness | implicit | unattached | **SENTENCE TOTAL** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | -0.375 | -1 | **-1.375** |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.75 | -0.75 | **0** |
| 2 | 0.125 | 0.125 | 0 | 0.375 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | -0.25 | **0.375** |
| 3 | -0.125 | 0 | -0.25 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.375 | **0** |
| 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.625 | **0.625** |
| 5 | 0 | 0.125 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | **1.125** |
| 6 | 0 | 0.5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | -0.1875 | 0.375 | **0.6875** |
| 7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.4375 | -0.25 | **0.1875** |
| 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | -0.25 | 1.5 | **1.25** |
| 9 | 0 | 0 | 0 | 0 | 0 | 0 | -0.125 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.75 | 0 | **0.625** |
| 10 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.65 | **0.65** |
| 11 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.225 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.8125 | **1.0375** |
| 12 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.375 | **0.375** |
| 13 | 0.625 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.875 | **1.5** |
| 14 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.5 | 0.75 | **1.25** |
| 15 | 0.375 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1.5 | **1.875** |
| **TOTAL** | **1** | **0.75** | **-0.25** | **0.375** | **0** | **0** | **-0.125** | **0** | **0** | **0.225** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **1.625** | **6.5875** | **10.1875** |

**Figure 4.2:** Example of output sheet detailing all aspect scores across all sentences of a review. Note that the leftmost column contains the sentence indices.

# 5

# Discussion

In this chapter, the results from the previous section are analysed. This includes a straightforward comparison of numerical metric results, and a more qualitative assessment of their performances, with analysis of error results to provide insight into the systems' shortcomings.

## 5.1 Research questions

The results of the project are examined below within the context of the four research questions stated in Section 1.2.

### 5.1.1 Rule-based system assessment

*How does a rule-based approach to sentiment analysis match compare to methods such as a neural network of simple count-based baseline?*

Examining the metrics outlined in Table 4.2, the most apparent conclusions are as follows:

1. The simplistic system, which merely counts sentiment terms across a review with no consideration for linguistic rules, matches or even marginally exceeds the performance of the rule-based system, though when binary categorisation metrics are factored in, it performs slightly worse.

2. When fine-tuned with only a small training set, BERT's performance within this domain is poor. Technically it achieves a perfect Recall rate of 1.0, but this is only because its predictions were entirely negative.

3. But when fine-tuned with a much more substantial training set, BERT provides the best performance of all, achieving a MAPE of 16.6%, RMSE of 1.239, Spearman's rho of 0.497 (ignoring the selectively masked set, which is covered below in Section 5.1.4) and F-score of 0.285.

All of this indicates a number of potential conclusions. The simplistic system is very straightforward to construct and easy to understand, and manages surprisingly good results for such a blunt approach. The rule-based system required a substantially greater investment of time and effort to reach the same sort of metric results.

However, it is also important to note that this rule-based system saw significant improvement over time and different iterations, as additional components were added to its process and its functionality was refined. Below are metrics collected across the development of the rule-based system. This development was conducted utilising the training set of 90 reviews, and so is not necessarily directly comparable to the later runs on the full test set of 10,000. It should also be noted that along with the major additions marked out at each stage, more minor, incremental changes were being made (such as further additions to SWN 3.0.666). These metrics should be taken more as indications of the gradual improvement of the system rather than perfect measurements of what each specific addition yields.

**Table 5.1:** Metrics of rule-based system across development period.

| | **MAPE** | **RMSE** | **Spearman's rho** |
|---|---|---|---|
| **Initial form** | 45.7% | 2.478 | 0.018 |
| **Contextual polarity added** | 45.3% | 2.387 | 0.084 |
| **Implicit aspect added** | 43.5% | 2.162 | 0.297 |
| **Unconnected sentiment added** | 50.4% | 2.287 | 0.367 |

It could thus be asserted that while the simplistic approach offers a quick method to achieve decent results, there may be greater potential available within a rule-based system if more time is taken to refine it further. The upper bounds of such a system's potential may not have been reached, while the simplistic system is, by definition, basic and allowing for little expansion without losing its intrinsic simplicity.

BERT's performance is clearly heavily reliant on its training. With a very small training set, it performs poorly. When trained on a bigger corpus, its performance surpasses that of either other system noticeably. While this does indicate that the deep learning approach can offer the best results, the rule-based system's lack of a need for a large training set can be taken as a point in its favour. Some texts are needed to examine and extract the rules that such a system will be based on, but the human reader's ability to more naturally find these patterns means they do not require tens of thousands of texts to come to their conclusions.

The results may also be compared to certain trivial baselines: as noted in Section 3.1, the mean rating across the test set is 7.6, while the median and mode are 8.0, and 93% of all reviews are positive if we use the binary categorical split where >5.0 = *positive* and 5.1< = *negative*. If a system always predicted the mean or mode/median (and, in turn, always predicted a positive categorical result), it would achieve the following metrics (note that Spearman's rho correlation coefficient cannot be calculated with a constant input):

**Table 5.2:** Trivial baseline metrics.

| | **MAPE** | **RMSE** | **Binary categorical accuracy** |
|---|---|---|---|
| **Always-mean (7.6, positive)** | 19.9% | 1.420 | 93% |
| **Always-median/mode (8.0, positive)** | 20.6% | 1.477 | 93% |

This achieves a greater MAPE and RMSE than the rule-based and simplistic systems, though not as great as BERT with the large training set. The accuracy is high, but this is due to the imbalance of the test set and corpus.

Below we examine some specific examples taken from cases where one of the three systems (rule-based, simplistic, BERT) made an entirely accurate prediction (with a 0.0 difference between prediction and actual rating), while the other two were inaccurate.

The rule-based system seems to perform best when points of sentiment are expressed in relation to the aspects the system searches for, as seen in Review 21853, which the rule-based system predicts accurately (5.9), while the others over-estimate its rating (9.5 from simplistic, 6.5 from BERT). In this, many sentiment expressions are indeed tied directly to aspects ("*the riffs are decent*", "*the largest problem with the song*", "*I must also compliment drummer Matt Swistak on his performance*", "*his lyrics are unintelligible*"), which may allow the rule-based system to make a more accurate prediction, as it specifically focuses on these for its calculations.

In Review 14510, the rule-based system again makes an accurate prediction (4.5), while the simplistic system under-predicts (3.2) and BERT over-predicts (6.9). This review utilises a lot of negative terms unattached to aspects. Both the simplistic and rule-based systems would pick up on, but these will be given less weight within the rule-based system, where the explicit and implicit aspects can help to off-set this. There's a lot of praise for the artist in question as a vocalist, for example ("*Ms. Deva is an accomplished and prolific vocalist, who's done backing vocals for many well known and respected metal acts*"). While both systems will detect these positive words, in the rule-based system, these lead to a positive (+2.0) score for the vocals as an aspect, which will count for more than simply detecting this as a single piece of unattached sentiment, as the simplistic system treats all terms found. This review also praises the artist outside of this particular album ("*Deva has made a career as a secondary vocalist in many bands, and that's just where she should stay. Her limited range fits well in such a role. Putting herself into the spotlight was simply a misstep.*"), and it's possible that BERT interprets this as positivity about this album itself.

In Review 12908, the simplistic system performs best as the album is filled with praise, which seems very easy to pick up on: "*Ninnghizhidda are one of the most original and vicious death metal bands working today, with a symphonic and technical touch that raises them above the ordinary or simply brutal*", "*"Blasphemy" is just as vicious and heavy as its successor, and every bit as cool and memorable.*", etc. However, in this review we see the reverse of in 14510: there, the praise assigned to the "vocals" aspect helped to off-set the negativity across the review. Here, an early line refers to another band entirely ("*I find it incomprehensible that crappy bands like Soilwork...*"). While both the simplistic and rule-based systems detect the negative presence of "*crappy*", in the rule-based system the presence of the term "*band*" assigns this sentiment to the implicit aspect, which is weighted quite heavily, while in the simplistic system it is just one of numerous sentiment terms all weighted equally.

Thus, it would seem the two systems' methods for making their calculations can be an advantage or disadvantage, depending on the context. Within the simplistic system, all sentiment terms are treated equally (in terms of their weight in the final calculations), so an erroneous one such as a term referring to another band will not weigh too heavily. In effect, the results are averaged-out somewhat.

With the rule-based system, the effect of sentiment terms (both those accurately assessed and those that represent an error like *crappy* in the above example) will depend upon how the system reads them, as this will alter their weight and impact. In a review where vital points of sentiment are expressed specifically in relation to implicit or explicit aspects, the rule-based system assigns these greater weight than they would be within the simplistic system, which may be closer to what the reviewer intends. But this greater weight will also apply to any erroneous terms picked up and assigned to an implicit or explicit aspect, leading to a greater detrimental effect. This indicates that the rule-based system is more vulnerable to this particular issue: an erroneous result in the simplistic system will likely be outweighed by the accurate result. Such a result in the rule-based system may be counted for more, but so too will correctly identified sentiment expressions. As such, the rule-based system is particularly dependent upon accurate identification of relevant sentiment expressions.

An additional point of interest can be seen in the most accurate predictions by the simplistic system: in total across the test set, the simplistic system made 211 perfect predictions. Of these, 80 (38%) were for 10/10 score reviews, representing a far greater percentage than such reviews hold within the full test set (where they make up only 2.6%). This could indicate that extremely positive reviews (at least within this domain) tend to use very straightforward language, or repeat it a lot, allowing the simplistic system to easily make a very strong prediction. This trend of the simplistic system is not so pronounced with negative reviews however: reviews of less than 5.1/10 make up 7% of the test set, but only 3.8% of the simplistic system's perfect predictions. This may indicate that negative reviews make greater use of methods of speech that the simplistic system does not pick up on, such as irony and sarcasm.

While BERT's results are harder to interpret, we can make some inferences from instances where BERT predicts accurately and the other two systems predict inaccurately. In Reviews 29047 and 28698, this was the case. These sort of cases seem likely due to problems with SWN, WSD or the parsing, as these are common tools shared by both the simplistic and rule-based systems, and in the reviews we see a number of words assigned negativity which should not be considered as such (bolded words below are considered negative within SWN, when within this context and/or domain, they should not be):

- *During the "thrash **attack** revival" of 2007, once former underground stalwarts such as Demolition **Hammer** and Morbid Saint have been basically re-discovered by a whole new generation.*
- *At **loud** volumes this album sounds like a warzone with the **double** bass sounding like **actual** machine gun fire.*
- *I wryly smile and **imagine** those 500 devotees, assuming all of the **early** adopters have been retained.*
- *On their followup EP In **Contemptuous** Defiance German Fiat Nox continues their thoughtful and convoluted, death tinged black metal, although maybe due to the EP **shorter** duration the Germans have expensed **away** with all **matters** of atmospheric, intros and acoustic interludes.*
- *Amok Hymn continues with zigzagy guitars flying around, **blasting** everything into prolific **chaos***
- *The return to more complex Fiat Nox appears to be **happening** with Unheiligkeitsklage, where **an unexpected** melody emerges with the **wintry** tremolo.*

Thus, both the simplistic and rule-based systems appear dependent upon the reliability of the tools they make use of to draw their conclusions.

### 5.1.2 Domain specificity

*Do domain-specific modifications improve the performance of sentiment analysis within this domain?*

Across all metrics, performance decreased when utilising the original, unmodified version of SWN (and with no intercept layer for new words). The MAPE increased from 24.2% to 28.7%, the RMSE increased from 1.993 to 2.411, the Spearman's rho correlation decreased from 0.314 to 0.238, and the F-score decreased from 0.250 to 0.193. As such, it would seem clear that domain-specific modifications in this instance have notably improved performance, even when based only on a small training set of 90 reviews.

For illustration, we can again examine specific examples, this time where SWN 3.0.666 produced much more accurate results. Of these, five in particular had perfect predictions with SWN 3.0.666, while the predictions using the original version of SWN were off by 1.8 or more. These five reviews have the IDs 129, 20101, 27258, 27817 and 27919, and below are some example sentences in which the bold words were added or modified in SWN 3.0.666 and led to a more accurate prediction (typically by removing negativity that is not intended within this domain):

- *...the **chilling** cues and **haunting** drones mesh perfectly with the elegantly evil images and go a long way in driving every scare deep into the bones.*

- *Just look at that cover. "**Scary** Muzak" is an appropriate descriptor for these creepy mood pieces.*
- *The **thrashy**, punk, and **black** metal group Cape of Bats released their album at the perfect time with their debut **Violent** Occultism.*
- *...the music concentrates on a **wicked** and erratic tone, boiling with an **evil** hate.*
- *One of the best tracks, Lord of Shadows, is an **angry** tune, aggressively punching out melodic **thrashy** riffs that have a muddy **black** metal sound to them...*
- *The guitars really meld well together here in a great **thrashy** and harmonious manner, but also delivering a great **wailing** solo towards the end.*
- *In terms of atmosphere The Mezmerist is squarely in the mid-80s, professing the then popular brand of **psychedelic doom** metal...*
- *A truly good **doom** album is a sea of **sadness**, a wide, deep, morass of **gloom** that the listener can plunge into and become lost.*
- *...the four tracks here each tell a different story of horrible **death** and endless **haunting**...*
- *...but all else is **crushing** riffs and engulfing drones, the useless struggle against an inevitable death. The end is especially good, **death**-rattle **growls** atop a **mournful** yet tuneful melody as the vocals turn clean and the track becomes grandiose.*
- *...continues this atmosphere with a melody riding atop the pounding **doom** riffs and throaty **growls**...*

### 5.1.3 Rationale outputs

*Can a rule-based system be made to output strongly polar sentences as rationales for its decision, and do these accurately represent some of the most important points of a review?*

The rule-based system outputs the 3 sentences with the greatest weight (according to its process) from each review as rationales justifying its predicted rating.

In order to assess the rationales output by the rule-based system, the top 50 outputs (as outlined in Section 4.2.1) were assessed. For each of these, the review text and the output rationales (i.e. the three sentences that the rule-based system considered to have the greatest sentiment weight) were examined, and it was judged whether the rationales accurately reflected both the text of the review and the predicted rating given by the system.

Of these, 26 were considered good reflections of the content of the review and predicted rating, 19 were considered more mixed in their reflections of the review content and predicted rating, and 6 were considered poor reflections of the content of the review and predicted rating. The rationales tended to match up in terms of their reflection of the review content and predicted rating, which is unsurprising since these are reviews where the predicted rating matched the actual rating.

In the instances where the rationales were weak or mixed as reflections of the review, this tended to be due to one of several common reasons.

Some reviews go through the album song by song, describing each, and while these can be meaningful for the overall rating of the review, they tend to be less than more summative sentences. These are sometimes returned as rationales.

Examples:

- "*The Night of Obliteration goes through several drastic changes with nice drum fills and speedy picking, all in just over 5 minutes.*"

This tells us something about the song in question, but not much about the wider album.

- "*Darker Times does have some guitar soloing, but the overall melody is grating and difficult to follow, and the chorus isn't much better.*"

As in the previous example, this isn't so informative about the full album.

Context is a recurring issue with some rationales: the rationale outputs are simply sentences from the review, and some sentences may rely on others to make sense, typically the preceding sentence. In these cases, the sentence may make little sense as an isolated rationale.

Examples:

- "*Partly, this is basically due to those extra guttural - as well as an extra low vomits of I.S.K.H.*"

Here we don't know what "this" refers to. This problem could in theory be addressed by implementing a pronoun resolution method.

- "*The sound was better, the performances better and, most importantly, the songwriting was much better.*"

This sentence is comparing an album to the band's past works, but we don't know that without the rest of the review.

In some cases, one or more rationales may be more descriptive of genre than quality. These can still provide useful information about the release, but may not mean as much in terms of the rating of the album. This is often the case if the review itself spends a significant amount of time describing the genre. Similarly, some rationales extract parts of the review more rooted in describing the band's history than the quality of this particular album.

Examples:

- "*My instincts proved correct as there are only flashes of Black Metal on Stronger Than Frost, in the form of occasional blast beats and rare raspy vocals.*"

This can be helpful information about the album's style, but isn't so meaningful when it comes to the rating.

- "*Deathwomb Catechesis is Pseudogod's debut full-length studio album under Finnish underground label Primitive Reaction and it is a pitch-black, frenzied whirlwind of both Black and Death Metal that these Russian fellows serve by catapults straight on our necks.*"

This sentence talks about the band's history and genre, but again doesn't say a lot relating to the rating.

### 5.1.4 Selective masking results

*Can these rationale outputs be used like a diagnostic tool for a neural network to help provide greater insight into its process?*

Ultimately, the use of selective masking on the test set did not impact the results of BERT very much. It increased the MAPE by 0.1 and the RMSE by 0.015, improved the Spearman's rho value by 0.002 and left the F-score unchanged. This would seem to suggest that BERT is making its decisions about sentiment in some fundamentally different way to the methods of the rule-based system: the rule-based system is highlighting these sentences as the most important by its measures, and then removing them has little impact on BERT, indicating that they are not so important to it. This suggests that either BERT bases its decisions on something else, or that it is more resilient to the inputs being cut down while still being able to make accurate predictions.

In order to seek greater insight, we will examine some specific examples where the prediction changed most dramatically between selectively masked and not.

In many cases where the selectively masked result differed significantly, this was because the original review was quite short, and thus masking 3 sentences removed a lot of its content, and left BERT with less to work with. Review 9813 for example, where the actual rating is 5.0 and BERT's prediction shifted from 4.2 to 8.2 (with masking), was reduced to:

> *I've just never been able to get that into Gun Barrel.*

In Review 6629, the actual rating is 7.5 and BERT's prediction shifted from 7.5 to 3.5. In this case, a little more text was left, but it included phrases that could easily be misinterpreted as negative (in bold):

> *Oh yeah. It's not original, you've heard it before, and you just won't care because you'll be busy scrapping yourself off the floor after being run over by this insane speedfest. It's just the perfect remedy when you've had a bad day and need to vent your frustration.*

There were some longer examples however where the selective masking significantly altered BERT's prediction, even when the remaining review text was of a decent length. In Review 3770 the actual rating is 7.5, and BERT's prediction shifted from

7.7 to 3.9. Looking at the sentences that were masked:

> *This is the second full length of this Hardcore/Thrash/Groove metal band from USA, this album happens to be a concept dragged from their debut album called "Flesh of the weak" where we can find a song called "Twelve", tune about a serial killer who hears 12 different demonic voices in his head, each voice tells him to kill in a different way and that's about it.*
>
> *The music is overall really fucking catchy, and I mean catchy as crazy glue or something, the beats have this constant groove which makes us head-bang over and over again and the riffs are structured in a meticulous way so we won't forget about them in a long time.*
>
> *The general feeling lines up in hardcore overtones, thrash beats here and there, a couple of samples (they overuse the "evil laughs" badly though) and a lot of groove, so yes this is a rough mix, but easy on the ear none the less.*

The shift in prediction alone indicates that BERT values the above sentences quite highly in its prediction. Interestingly, while none of them are truly negative, only the second is noticeably positive. The others are more descriptive of the album's sound and style. This could indicate that BERT is drawing a connection between descriptive passages like that and a positive review. Even if such sentences are not positive in isolation, it could reveal a pattern that BERT has learned during training. In an image-recognition task, a neural network may pick up on contextual clues (like the presence of a road to identify images of cars) that are not themselves part of the sought object, but are indicative of its presence. Similarly, it could be that these descriptive passages are commonplace in positive reviews.

In Review 21440, BERT's prediction actually improved with selective masking: the actual rating is 7.0, and BERT's original prediction was 4.1, but with masking it became 8.0. This could be because BERT was considering certain parts in the masked sentences to be negative, harming its original prediction (bold):

> *In fact, lots of great bands have had a **weak album** in their discography.*
>
> *The way they weave together those beautiful atmospheric melodies and spread feelings through their songs make Artrosis its raison d'etre.*
>
> *Sometimes **I would like to hear** some male vocals to give this album a greater metal edge.*

## 5.2 Additional research questions

In addition to the results assessed above, some results can also provide further insights into areas not explicitly laid out among the original research questions.

These could be considered additional research questions that posed themselves over the course of the project.

### 5.2.1 Discourse analysis (sentiment flow)

For these results, a research question such as the following could be considered: "*Does the flow of sentiment across reviews, as assessed by the rule-based system, highlight meaningful patterns within these reviews?*"

As outlined in Section 4.2.1, the sentence sentiment scores across each of the 50 best results from the rule-based system were plotted on a graph, to see if they revealed any common patterns or trends in the flow of the sentiment across reviews.

However, these plots do not appear to show any common patterns that could be considered reliable enough to base additional rules on. They were examined manually and in different groupings (e.g. short reviews vs. long reviews, positive ratings vs. negative ratings), but did not appear to display meaningful trends. It is possible that with a more mathematically robust method of examination there may be some subtle underlying trends. The rule-based system itself is also not perfect, and it is possible that if its outputs were improved, these plots may become more meaningful.

### 5.2.2 Explicit aspect polarity results

For these results, a research question such as the following could be considered: "*How accurately can a rule-based system score individual aspects within these reviews?*"

For this, the same subset of 50 top results from the rule-based system were used. Each of these was manually annotated for explicit aspect polarity. For simplicity's sake, this was done on a binary division, marking out positive and negative sentiment expressions within each review, and only evaluated based on the explicit aspects, not implicit or unconnected sentiment. In each case, this could then be compared to the output of the rule-based system for these 50 reviews to see how much it agreed with the manual annotation.

In total across the 50 reviews, 189 aspects were marked during manual annotation (128 positive and 61 negative). Note that this annotation does not account for the degree of sentiment, nor multiple expressions about a single aspect: it is intended to see if the rule-based system can accurately reflect human detection of sentiment regarding these possible aspects. Across these 50, the rule-based system highlighted 151 aspects (119 positive and 32 negative).

Considering the manual annotation as a ground truth value in this instance, we can treat this as a multi-class classification with three possible classes for each aspect in each review: no sentiment, positive sentiment or negative sentiment (in some instances it is possible the system detected sentiment about an aspect, but that it balanced back out to 0 overall, so in this context it may be more accurate to describe no sentiment as neutral sentiment instead). From this, we can create a confusion matrix:

**Table 5.3:** Confusion matrix treating explicit aspect polarity assessment as a multi-class classification.

| | **System detects no/neutral sentiment** | **System detects negative sentiment** | **System detects positive sentiment** |
|---|---|---|---|
| **Manual detects no/neutral sentiment** | 653 | 17 | 42 |
| **Manual detects negative sentiment** | 34 | 11 | 15 |
| **Manual detects positive sentiment** | 62 | 4 | 62 |

This yields the following True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) rates for each of the three classes:

**Table 5.4:** TP/TN/FP/FN of explicit aspect sentiment polarity.

| | **TP** | **TN** | **FP** | **FN** |
|---|---|---|---|---|
| **No/neutral sentiment** | 653 | 92 | 95 | 59 |
| **Negative sentiment** | 11 | 819 | 21 | 49 |
| **Positive sentiment** | 62 | 715 | 57 | 66 |

This in turn can be used to calculate Precision, Recall and F-scores for each category.

**Table 5.5:** Metrics of explicit aspect sentiment polarity.

| | **Precision** | **Recall** | **F-score** |
|---|---|---|---|
| **No/neutral sentiment** | 0.87 | 0.92 | 0.89 |
| **Negative sentiment** | 0.34 | 0.18 | 0.24 |
| **Positive sentiment** | 0.52 | 0.48 | 0.50 |

The above seems to indicate that the rule-based system is better at accurately detecting positive sentiment than negative, and that in particular it could benefit from improved means of detecting negative sentiment when it is present.

## 5.3 Error analysis

In addition to the 50 best results from the rule-based system utilised for analysis above, the 50 worst results were also extracted (i.e. those with the largest gap between the predicted rating and the ground truth value). These are utilised below for some detailed error analysis, to try to pinpoint the flaws in the system.

From examination of these reviews, a number of common trends could be identified as causes for the rule-based system's inaccurate predictions:

- The system will often incorrectly extract sentiment from **album titles**, **song titles** and **band names**. In review 20195, ***Slayer***, ***Morbid Angel*** and ***Raining*** *Dead Angels* all generated negative sentiment scores, as did *In the Sign of* ***Evil***, ***Outbreak*** *of* ***Evil***, *Burst Command til* ***War***, *Obsessed by* ***Cruelty*** and *Persecution* ***Mania*** in Review 15227. In each of the examples above, the relevant sentiment term is marked in bold. Naturally, such instances should not convey sentiment alone. It is possible that reliable tagging as proper nouns (and excluding proper nouns from sentiment scores) could mitigate this issue.

- The system struggles with **irony and sarcasm**. This is particularly evident in the highly negative Review 22140, where phrases such as *The members have such great names*, *this is hilarity to the fullest coming from them*, *such awesome lyrical prose*, and *this masterpiece* all contribute positivity to the final score, when they actually express negative sentiment.

- The **Word Sense Disambiguation** commonly led to issues with neutral terms yielding sentiment. The word like, for example, is often used simply to compare two things, but was commonly misinterpreted as referring to positive preference. This also caused issue when a relevant entry had been modified in SWN 3.0.666, but an incorrect (unmodified) one was pointed to by the disambiguation function. SWN 3.0.666 had been modified, for example, to account for the term *sadness* often (within this domain) not actually being an indicator of negativity. But in Review 20195, a different, unmodified definition was used, and so the sentence, *Without skipping a beat the album can then switch into acoustic bittersweet sadness* yielded incorrect negative sentiment to the overall rating.

- While the changes made to **SWN** did improve performance (see Section 5.1.2), further work is still needed there. Terms such as attack, assault and torment require modification, as they will often not truly impart negativity in this domain. There are also numerous words that impart sentiment when they are more commonly neutral (e.g. *do*, *early*, *comparison*, *think*, *a*, *view*, *double*).

- In some reviews, sentiment was detected when the reviewer was in fact referring to **other bands or previous albums** by the band in the review. Such references should not impart sentiment onto the album being currently reviewed.

- The system struggles with **negation**. This is a known issue within NLP [6] [14], and while some steps were taken on it within the rule-based system (see Section 3.2.4.1), further work could improve on this.

- In some cases, the issues lies in part with **the review text itself**: Review

14755, for example, can be misleading because the entire first half of the review is spent outlining what the reviewer wanted the album to sound like, and so this leads to a much more positive predicted rating. In Review 15227, the opening paragraph is highly negative, but this is referring to dissatisfaction with the record label's handling of the album, and does not actually reflect the reviewer's feelings on the album itself, which was given a 10/10 score. It is plausible that even human annotators may differ in their assessment of some such examples.

# 6
# Conclusion

In this chapter I discuss the over-arching conclusions I believe can be drawn about the project based on its results. This section also includes discussion of problems or limitations that were identified about the project and hypothetical future avenues for expansion or refinement.

## 6.1 Primary conclusions

In summary, while the rule-based system constructed within this thesis project showed definite improvement over the course of its development, it took a significant amount of time and work to even meet the level of a simplistic system. BERT demands a substantial amount of data to be properly fine-tuned, but once this is done, it seems readily able to exceed the performance of either of the other systems.

However, the work done with the rule-based system is not entirely conclusive, and there remain areas of improvement that could be made to it. The system benefitted from fine-tuning SWN to be more domain specific, which itself could be continued further. The rationales output by the rule-based system also show some promise as potential summaries of a review's important points, and could hypothetically be utilised to better filter reviews and make recommendations by an individual's preferences, but more work would be needed to further refine these, either for their own sake, or for use as a diagnostic tool for neural networks.

Nevertheless, based on the work done and progress made thus far, I do believe there is potential to build upon the system and push its capabilities further, especially within this domain. Below are summarised some of the known issues with the system, and some thoughts for future potential development.

## 6.2 Known problems & limitations

As detailed in the error analysis in Section 5.3, there are a number of known problems and limitations of the rule-based system constructed within this project. Some of these represent known areas in which there is room for improvement, others are issues highlighted during development and analysis of the system's outputs.

Names of bands, albums, songs and genres can all cause unintended sentiment. This is especially an issue within the hard rock/heavy metal domain, where many titles may be rooted in conventionally negative topics (e.g. the death metal genre, bands names such as My Dying Bride or Pain of Salvation, misanthropic or blasphemous lyrical themes, etc). In addition to the need to avoid incorrect sentiment detection in this manner, accurate recognition of band, album and song entities could itself be a useful feature for a domain-tuned rule-based system. If such recognition were possible, for example, a system could more easily read a review and highlight praised or disliked songs.

The system's performance is also inherently tied to the tools that it is built with. Shortcomings or flaws with these will impact on the resulting system. In some cases, these tools can be directly tuned for this task, as SWN was when modified for this project. In other cases, there may be alternative options that were not tried in this project, but could provide more useful results, such as alternative parsers or word sense disambiguation tools.

Negation is a specific point within the rule-based system which could have benefitted from further work, as illustrated by some of the mis-classifications seen in the error analysis. A simple handling was implemented alongside the system's handling of modifier terms, but in truth negation requires significantly more time devoted to it to properly address it.

Some difficulties can also be caused due to issues within the review texts themselves. Spelling mistakes may lead to mistakes in word dependencies or sentiment values, as can incorrect punctuation use (lack of full stops can make assessment by-sentence difficult, for instance). The system of a reviewer assigning a numerical rating and writing a review text to match is also inherently an imprecise one: a review text may read especially negative or positive, but this may not be reflected in the score assigned, and each individual may have their own definitions of what constitutes a positive or negative score. To some, 6/10 may indicate something generally positive, while to others it may represent more significant flaws. In some cases, the rating may not even attempt to be an accurate reflection, as can be seen in Review 777's opening:

> *By popülar demand, I am rating my new review. By ünpopülar demand, the rating has nothing to do with the albüm's qüality, büt with it's content: Evil! Speed! Darkness! Steel! Satan! Goats! 666!*

In addition, the manual annotations performed on some reviews could not be conducted across the full corpus for reasons of time. Not only would it be beneficial to perform annotations on a wider scale, but conducting them with only one individual leaves them open to greater subjectivity. Conducting annotations with multiple individuals could provide more reliable results.

## 6.3 Future work

Were the work of this project to continue further, there are numerous potential avenues to explore in seeking to improve the results:

1. As mentioned above, recognition of relevant entities such as album, band and song names, and band members, could be very useful for this domain in better recognising the target of sentiment expressions (including when the target is something outside of the album being reviewed, such as a band's past albums). In a similar vein, reviews of this domain will often make comparisons to existing bands and works (e.g. *the riffs on this album would do Iron Maiden proud*), and recognition of these could also be useful.

2. Expansion of manual annotation of aspects would also be useful, and could lead to further exploration of how the aspects so commonly referred to in these reviews could be employed in other ways to analyse the reviews.

3. Also mentioned previously, other options for lemmatisation, parsing and word sense disambiguation could all be explored, and may offer new options or superior results.

4. The removal of a small segment at the start of a review as an introduction produced some interesting results (see Section 3.4), which indicated that such an introduction could be removed without harming predictions, or even potentially improving them. Through closer examination of these introductory portions of reviews, the manner in which they are defined by my system could potentially be refined and improved, to better ensure that irrelevant or misleading text is removed, and that nothing relevant is inadvertently lost.

5. Aside from general work to try to improve the system and the rationales it outputs, further work could be done in trying to turn these rationale outputs into meaningful summarisations of the reviews. This would likely take substantial work, first to ensure that the most meaningful sentences are used as a basis, and secondly to transform these into a readable and valuable summary output.

6. The topic of genre specificity also yielded interesting results, and could be explored further by delving into specific subgenres or styles. Certain words may more commonly be used positively/negatively in some subgenres than others, for example, and so models trained on a particular subset may yield better results when deployed on a new text of a similar subgenre. Just as the domain specificity work done so far improved results, it could be that pushing this further still may yield better results still.

7. There are also some additional test set runs that would have been run had there been sufficient time available. Running BERT on another corpus entirely could have revealed more about its own handling of different domains,

and masking not only certain sentences but all sentiment terms within a review could have indicated how it values and handles sentiment terms in general.

8. Lastly, the differing advantages and disadvantages of the systems compared here gives rise to the idea of combining them to try to make use of the best features of each. The simplistic and rule-based systems, for example, could hypothetically be combined for making joint predictions, such as the simplistic system making a broad assessment which could be fine-tuned by the rule-based system.

# Bibliography

[1] Fang, X. & Zhan, J. (2015). Sentiment Analysis Using Product Review Data. *Journal of Big Data* (2, 5).

[2] Singh, V., Piryani, R., Uddin, A. & Waila, P. (2013). Sentiment analysis of movie reviews: a new feature-based heuristic for aspect-level sentiment classification. *2013 International Mutli-Conference on Automation, Computing, Communication, Control and Compressed Sensing (iMac4s)*. pp. 712-717.

[3] Nasukawa, T. & Yi, J. (2003). Sentiment Analysis: Capturing Favorability Using Natural Language Processing. *K-CAP '03: Proceedings of the 2nd International Conference on Knowledge Capture*. pp. 70-77.

[4] Feldman, R. (2013). Techniques and Applications for Sentiment Analysis. *Communications of the ACM* (56, 4). pp. 82-89.

[5] Wilson, T., Wiebe, J. & Hoffmann, P. (2005). Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. *Proceedings of human language technology conference and conference on empirical methods in natural language processing.* pp. 347-354.

[6] Pang, B., Lee, L. & Vaithyanathan, S. (2002). Thumbs up? Sentiment Classification using Machine Learning Techniques. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. pp. 79-86.

[7] Medhat, W., Hassan, A. & Korashy, H. (2014). Sentiment analysis algorithms and applications: a survey. *Ain Shams Engineering Journal* (5, 4). pp. 1093-1113.

[8] Turing, A. (1950). I. - Computing Machinery and Intelligence. *Mind* (LIX, 236). pp. 433-460.

[9] McGuigan, B. (2021). *How Big is the Internet?*. EasyTechJunkie. Retrieved from: https://www.easytechjunkie.com/how-big-is-the-internet.htm

[10] Esuli, A. & Sebastiani, F. (2006). SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining. *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. pp. 417-422.

[11] Liu, X., He, P., Chen, W. & Gao, J. (2019). Improving Multi-Task Deep Neural Networks via Knowledge Distillation for Natural Language Understanding. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. pp. 4487-4496.

[12] DeYoung, J., Jain, S., Rajani, N., Lehman, E., Xiong, C., Socher, R. & Wallace, B. (2020). ERASER: A Benchmark to Evaluate Rationalized NLP Models. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. pp. 4443-4458.

[13] Yessenalina, A., Choi, Y. & Cardie, C. (2010). Automatically generating annotator rationales to improve sentiment classification. *Proceedings of the ACL 2010 Conference Short Papers.* pp. 336-341.

[14] Taboada, M., Brooke, J., Tofiloski, M., Voll, K. & Stede, M. (2011). Lexicon-based Methods for Sentiment Analysis. *Computational Linguistics, 37*(2), pp. 267-307.

[15] Taboada, M. (2016) Sentiment analysis: An overview from linguistics. *Annual Review of Linguistics, 2.* pp. 325-347.

[16] Thet, T. T., Na, J. C., Khoo, C. S., & Shakthikumar, S. (2009). Sentiment analysis of movie reviews on discussion boards using a linguistic approach. *Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion.* pp. 81-84.

[17] Meena, A., & Prabhakar, T. V. (2007). Sentence level sentiment analysis in the presence of conjuncts using linguistic analysis. *European conference on information retrieval.* pp. 573-580. Springer, Berlin, Heidelberg.

[18] Ke, P., Ji, H., Liu, S., Zhu, X., & Huang, M. (2019). SentiLARE: Sentiment-aware language representation learning with linguistic knowledge. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP).* pp. 6975-6988.

[19] Xia, R., & Zong, C. (2010). Exploring the use of word relation features for sentiment classification. *Coling 2010: Posters*, pp. 1336-1344.

[20] Baccianella, S., Esuli, A., Sebastiani, F. (2010). SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA). pp. 2200-2204.

[21] Benamara, F., Taboada, M., Mathieu, Y. (2017). Evaluative Language Beyond Bags of Words: Linguistic Insights and Computational Applications. *Computational Linguistics* (2017) 43 (1), pp. 201–264.

[22] Carenini, G., Ng, R. T., Zwart, E. (2005). Extracting Knowledge from Evaluative Text. *K-CAP '05: Proceedings of the 3rd international conference on Knowledge capture* (October 2005), pp. 11-18.

[23] Hunston, S. (2011). *Corpus Approaches to Evaluation: Phraseology and Evaluative Language.* Routledge, New York.

[24] Hayes, K. (2022). *Meta-Metal.* Google Sites. Retrieved from: https://sites.google.com/site/tymellsheavymetaluniverse/

[25] Stanford NLP Group (2018). Stanza - *A Python NLP Package for Many Human Languages.* Stanford University. Retrieved from: https://stanfordnlp.github.io/stanza/

[26] Princeton University (2010). *WordNet 3.0 database statistics.* Princeton University. Retrieved from: https://wordnet.princeton.edu/documentation/wnstats7wn

[27] NLTK Project (2009). *NLTK Word Sense Disambiguation module.* NLTK Project. Retrieved from: https://www.nltk.org/api/nltk.wsd.htmlmodule-nltk.wsd

[28] Sauri, R. (2008). *A Factuality Profiler for Eventualities in Text.* Ph.D. dissertation, Brandeis University, Waltham, MA.

[29] Pandas Development Team (2008). *pandas.DataFrame.sample.* Pandas Development Team. Retrieved from: https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.sample.html

[30] Taboada, M., Brooke, J., Tofiloski, M., Voll, K. & Stede, M. (2011). *SO-CAL dictionaries: int_dictionary1.11.txt.* GitHub. Retrieved from: https://github.com/sfu-discourse-lab/SO-CAL/tree/master/Resources/dictionaries/English

[31] Goldberg, Y. (2017). *Neural Network Methods for Natural Language Processing.* Morgan & Claypool.

[32] Ashraf, S. (2021). *Basic Concepts of Neural Network.* Sciencious. Retrieved from: https://sciencious.com/basic-concepts-of-neural-network/

[33] Devlin, J., Chang, M. W., Lee, K. & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers).* pp. 4171–4186.

[34] bert-base-cased, Hugging Face (accessed May 2022). Retrieved from: https://huggingface.co/bert-base-cased

[35] Metal Crypt (n.d.). Retrieved from: http://www.metalcrypt.com/

[36] Metal Reviews (n.d.). Retrieved from: https://www.metalreviews.com/

[37] Metal Storm (n.d.). Retrieved from: https://metalstorm.net/home/

[38] Metal Underground (n.d.). Retrieved from: http://www.metalunderground.com/

[39] Universal Dependencies (2015). Retrieved from: https://universaldependencies.org/

[40] Mozetič I, Grčar M, Smailović J (2016). Multilingual Twitter Sentiment Classification: The Role of Human Annotators. *PLoS ONE* 11(5): e0155036. Retrieved from: https://doi.org/10.1371/journal.pone.0155036

# A

# Appendix 1

## A.1 A.1 SWN 3.0.666

Below are the details of the amendments made to SWN to create SWN 3.0.666. This includes new entries (captured in the *intercept layer* within the Python code) and entries for which the sentiment scores were modified.

### A.1.1 New entries

| New word(s) | POS | ID | PosScore | NegScore | SynsetTerms | Gloss |
|---|---|---|---|---|---|---|
| Hard-rocking/<br>Headbanging | a | 66600001 | 0.375 | 0 | hard-rocking#1<br>headbanging#1 | playing hard rock or heavy metal music in an effective manner,<br>"the album is a hard-rocking time" |
| Rock | v | 66600002 | 0.375 | 0 | rock#8 | to play rock music, typically to do so well,<br>"this band rocks hard" |
| Masterclass/<br>Masterwork | n | 66600003 | 0.625 | 0 | masterclass#1<br>masterwork#1 | highly impressive work,<br>"the album is a masterclass of heavy metal" |
| Brutaliser | n | 66600004 | 0.25 | 0 | brutaliser#1<br>brutalizer#1 | music that is especially heavy and effective,<br>"every song was a fearsome brutaliser of a track" |
| By-the-numbers/<br>Run-of-the-mill | a | 66600005 | 0 | 0.25 | by-the-numbers#1<br>run-of-the-mill#1<br>by_the_numbers#1<br>run_of_the_mill#1 | dull and predictable,<br>"the whole album was very by-the-numbers" |
| Cringe-worthy/<br>Cringeworthy/<br>Cringey/<br>Cringy | a | 66600006 | 0 | 0.75 | cringe-worthy#1<br>cringeworthy#1<br>cringey#1<br>cringy#1 | provokes an awkward or embarrassed reaction,<br>"the over-the-top lyrics were cringeworthy" |
| Wankery | n | 66600008 | 0 | 0.5 | wankery#1 | music with too much emphasis on technical showmanship,<br>to the detriment of the overall quality,<br>"the song is filled with too much flashy wankery" |
| Thrashy/<br>Thrash-y | a | 66600009 | 0.375 | 0 | thrashy#1 | music with qualities typical of thrash metal, usually done well,<br>"Slayer are a very thrashy band" |
| Catchiness | n | 66600010 | 0.25 | 0.125 | catchiness#1 | a measure of how memorable a piece of music is,<br>"the album's biggest strength is its catchiness" |
| Headbang | v | 66600013 | 0.375 | 0 | headbang#1<br>bang#7 | to nod one's head, a sign of approval of hard rock/heavy metal music,<br>"this music is perfect to headbang to" |
| Re-defining | a | 66600014 | 0.375 | 0 | re-defining#1 | something that is of sufficient quality to re-define its type,<br>"the music here is genre re-defining" |
| Skip-worthy | a | 66600015 | 0.25 | 0.5 | skip-worthy#2 | music that is considered weak enough to pass over,<br>"unfortunately, the third song on the album is totally skip-worthy" |
| Forgettability | n | 66600017 | 0 | 0.5 | forgettability#1 | a measure of how forgettable a piece of music is,<br>"the album's biggest weakness is its forgettability" |
| Rollercoaster | n | 66600020 | 0.375 | 0 | rollercoaster#1 | a ride, in this context intended to describe something exciting,<br>"this song is a real rollercoaster" |
| Cliche/<br>Cliche | a | 66600021 | 0 | 0.375 | cliche#2<br>cliche#2 | a description of something as overdone and predictable,<br>"the lyrics are so cliche they lack impact" |
| Trve/<br>Tr00/<br>Kvlt/<br>Cvlt | a | 66600023 | 0 | 0.5 | trve#1<br>tr00#1<br>kvlt#1<br>cvlt#1 | slang terms, corruptions of 'true' and 'cult',<br>intended to mock black metal that takes itself too seriously,<br>"this album tries so hard to be tr00 and kvlt" |
| Stratospheric | a | 66600024 | 0.25 | 0 | stratospheric#1 | very good, impressive,<br>"the music here is stratospheric" |
| Asshat | n | 66600025 | 0 | 0.625 | asshat#1 | an insulting term similar to 'asshole',<br>"the band just come across as asshats" |
| Knuckledragger | n | 66600027 | 0 | 0.375 | knuckledragger#1 | something unintelligent,<br>"every song is a knuckledragger dominated by dumb lyrics" |
| Clubcore | n | 66600028 | 0 | 0.5 | clubcore#1 | a derogatory term for music suited to a nightclub,<br>"the band just come off as making clubcore music" |
| Br00tal/<br>Brootal | a | 66600029 | 0 | 0.5 | br00tal#1<br>brootal#1 | slang terms, corruptions of "brutal",<br>intended to mock extreme metal that tries too hard to be heavy or aggressive,<br>"the band thinks they're so br00tal, it's sad" |
| Br00tality/<br>Brootality | n | 66600030 | 0 | 0.5 | br00tality#1<br>brootality#1 | slang terms, corruptions of "brutality",<br>intended to mock extreme metal that tries too hard to be heavy or aggressive,<br>"the br00tality flaunted here is just too much to take seriously" |
| Must-hear/<br>Must-listen | n | 66600031 | 0.375 | 0 | must-hear#1 | music that is very good and deserves to be listened to,<br>"this album is an absolute must-hear" |
| Poor-man's | a | 66600032 | 0 | 0.875 | poor-man's#1<br>poor_man's#1 | something inferior,<br>"this band is honestly just a poor-man's Iron Maiden" |
| Standout | n | 66600033 | 0.375 | 0 | standout#1 | something that stands out as especially good,<br>"the last song is the real standout of the album" |
| Standout | a | 66600034 | 0.375 | 0 | standout#2 | a description of something that stands out as especially good,<br>"this one's a real standout song" |
| Mallcore | n | 66600035 | 0 | 0.5 | mallcore#1 | a derogatory term for music seen as overly commercial and<br>intended to appeal to young people at shopping malls,<br>"this is just more mallcore" |
| Oversaturation | n | 66600036 | 0.125 | 0.5 | oversaturation#1 | when something is overly commonplace to the point where it loses impact,<br>"the oversaturation of power metal makes this feel weak" |
| Re-hash | n | 66600037 | 0 | 0.375 | re-hash#2 | something that treads familiar ground, typically seen as dull because of it,<br>"this is just a re-hash" |
| Re-hash | v | 66600038 | 0 | 0.375 | re-hash#1 | to tread familiar ground, typically producing something seen as dull,<br>"they just re-hashed their last album again" |

| New word(s) | POS | ID | PosScore | NegScore | SynsetTerms | Gloss |
|---|---|---|---|---|---|---|
| Butthurt | a | 66600039 | 0.125 | 0.75 | butthurt#1 | a mocking term describing someone who is upset at something,<br>"the band just got butthurt that no one liked their last album" |
| Burnout | n | 66600040 | 0.125 | 0.625 | burnout#1 | the feeling of being exhausted, usually about something in particular,<br>"the band is feeling burnout over this" |
| Fist-pumping/<br>Horn-throwing/<br>Neck-snapping | a | 66600043 | 0.375 | 0 | fist-pumping#1<br>horn-throwing#1<br>neck-snapping#1 | a description of hard rock/heavy metal music that encourages an enthusiastic response,<br>"the band make fist-pumping music sound easy" |
| Fist-pumper/<br>Horn-thrower/<br>Neck-snapper | n | 66600044 | 0.375 | 0 | fist-pumper#1<br>horn-thrower#1<br>neck-snapper#1 | a description of hard rock/heavy metal music that encourages an enthusiastic response,<br>"this album's a real neck-snapper" |
| Mudslinging | n | 66600045 | 0.5 | 0.125 | mudslinging#1 | casting aspersions or insults,<br>"the mudslinging present here is offensive" |
| Listenability | n | 66600046 | 0.625 | 0 | listenability#1 | a measure of how accessible and easy to listen to music is,<br>"the listenability of this album is outstanding" |
| Awesomeness | n | 66600047 | 0.875 | 0.125 | awesomeness#1 | a measure of how awesome something is,<br>"this new album is packed with awesomeness from top to bottom" |
| Thrashiness | n | 66600048 | 0.375 | 0 | thrashiness#1 | a measure of how much effective thrash metal content there is in music,<br>"the thrashiness in this song is great" |
| Doomster | n | 66600049 | 0.375 | 0 | doomster#1 | a descriptive term of someone who plays doom metal, usually doing so well,<br>"these guys are experienced doomsters" |
| No-holds-barred | a | 66600050 | 0.5 | 0.25 | no-holds-barred#1 | a descriptive term for music that holds nothing back,<br>"this stuff is a no-holds-barred party in musical form" |
| Snore-fest/<br>Bore-fest | n | 66600051 | 0 | 0.25 | snore-fest#1 bore-fest#1 | something that is very boring,<br>"this album is a total snore-fest" |
| Must-have | n | 66600052 | 0.375 | 0 | must-have#1 | something that is of high enough quality to be considered essential to own,<br>"this album is a must-have" |
| Vomit-inducing | a | 66600053 | 0 | 0.5 | vomit-inducing#1 | something that induces the listener to vomit,<br>"the music here is so bad it's vomit-inducing" |
| Turn-off | n | 66600054 | 0 | 0.75 | turn-off#1 | something that reduces the interest of the listener,<br>"the vocals here are the real turn-off" |
| Pitch-perfect | a | 66600055 | 1 | 0 | pitch-perfect#1 | something that is perfect,<br>"the riffs on this album are pitch-perfect" |
| Clunker | n | 66600056 | 0 | 0.625 | clunker#1 | a descriptive term for something ineffective,<br>"this song is the worst, it's a real clunker" |
| Rip-off/<br>Ripoff | n | 66600057 | 0 | 0.25 | rip-off#2 ripoff#1 | something that is considered to have plagiarised someone or something else,<br>"this album is a total rip-off of another band" |
| Rip-off/<br>Ripoff | v | 66600058 | 0 | 0.25 | rip-off#2 ripoff#1 | to plagiarise someone or something else,<br>"frankly, this band just ripped off their peers" |
| Pulse-racing | a | 66600059 | 0.375 | 0 | pulse-racing#1 | something exciting,<br>"the guitars on this are pulse-racing" |
| Trainwreck | n | 66600061 | 0 | 0.375 | trainwreck#1 | something disastrous,<br>"the album as a whole is a trainwreck" |
| Jaw-dropping | a | 66600066 | 0.375 | 0 | jaw-dropping#1 | something that makes someone's jaw drop in shock, usually at how good it is,<br>"this album is so good, it's jaw-dropping" |
| Jaw-droppingly | r | 66600067 | 0.375 | 0 | jaw-droppingly#1 | something that makes someone's jaw drop in shock, usually at how good it is,<br>"this album is jaw-droppingly good" |
| Infectiousness | n | 66600068 | 0.25 | 0.125 | infectiousness#1 | a measure of how memorable a piece of music is,<br>"the music here is truly infectious, it's so hard to get out of your head" |
| Crustiness | a | 66600069 | 0.375 | 0 | crustiness#1 | a measure of music that is particularly rough or gritty, usually in an intentional way,<br>"the crustiness that this album displays makes it a cut above the rest" |
| Self-parody | n | 66600071 | 0 | 0.375 | self-parody#1 | something that inadvertently parodies itself,<br>"the cliches are so prominent here, the album becomes a self-parody" |
| Pitfall/<br>Pit-fall | n | 66600072 | 0 | 0.625 | pitfall#9 pit-fall#1 | a danger,<br>"the pitfall with this kind of music is going on too long" |
| Sellout/<br>Sell-out | n | 66600073 | 0 | 0.625 | sellout#3 sell-out#2 | a person or product that is considered to sacrifice artistic integrity for commercial gain,<br>"the lead singer is a sell-out" |
| Sellout/<br>Sell-out | v | 66600074 | 0 | 0.625 | sellout#2 sell-out#1 | to sacrifice artistic integrity for commercial gain,<br>"they really sold-out on this album" |
| Skull-crushing | a | 66600075 | 0.25 | 0 | skull-crushing#1 | a description of heavy music that is particularly effective,<br>"this is proper skull-crushing death metal, as it should be" |
| Skull-crushingly | r | 66600076 | 0.25 | 0 | skull-crushingly#1 | a description of heavy music that is particularly effective,<br>"the music here is skull-crushingly good" |

| New word(s) | POS | ID | PosScore | NegScore | SynsetTerms | Gloss |
|---|---|---|---|---|---|---|
| Eargasm | n | 66600077 | 0.375 | 0 | eargasm#1 | a slang term for especially enjoyable music, "every track here is a total eargasm" |
| Hard-hitter | n | 66600078 | 0.25 | 0 | hard-hitter#1 | something that hits hard, is effective, "the opening song is a real hard-hitter" |
| Hard-hitting | a | 66600079 | 0.25 | 0 | hard-hitting#1 | something that hits hard, is effective, "the last song on here is so hard-hitting" |
| Earworm | n | 66600080 | 0.25 | 0.125 | earworm#1 | a slang term for something that is especially memorable or catchy, "every song on here is a real earworm" |
| Infectious | a | 66600081 | 0.25 | 0 | infectious#4 | a measure of how catchy or memorable a piece of music is, "this rhythms in this song are really infectious" |
| Heavy | a | 66600082 | 0.375 | 0 | heavy#28 | a descriptor of the intensity or prevelance of characteristics typical of heavy metal music, "Black Sabbath were much heavier than their contemporaries" |
| Heaviness | n | 66600083 | 0.375 | 0 | heaviness#6 | a measure of the intensity or prevelance of characteristics typical of heavy metal music, "the heaviness in this album is impossible to ignore" |
| Harsh | a | 66600084 | 0 | 0 | harsh#7 | a descriptor of rough, aggressive vocals commonly used in forms of extreme metal music, "there are lots of harsh vocals on the new Arch Enemy album" |
| Harshness | n | 66600085 | 0 | 0 | harshness#5 | a measure of how rough and aggressive a piece of music is, "there's a lot of harshness in this new song" |
| Filler | n | 66600086 | 0 | 0.25 | filler#6 | parts of a musical album that are considered weaker or less important, "there's far too much filler on this album" |
| Generic | a | 66600087 | 0 | 0.375 | generic#4 | a description of something as so typical of a genre that it becomes bland, "the music is just so generic" |
| Shredding | n | 66600088 | 0 | 0 | shredding#2 | a style of music emphasising technically proficient guitar playing, "the shredding on this album is amazing" |
| Shred | v | 66600089 | 0 | 0 | shred#2 | to play guitar in a technically proficient manner, "when this guitarist shreds, it's like nothing else" |
| Killer | a | 66600090 | 0.375 | 0 | killer#4 | a term for a satisfying piece of music, especially heavy music, "this album is just killer" |
| Solid | a | 66600091 | 0.25 | 0 | solid#16 | a positive description of enjoyable music, "every song is absolutely solid and worth listening to" |
| Distorted | a | 66600092 | 0 | 0 | distorted#3 | music that has had its sound output modified, usually by increasing the gain, "the guitars on here are really heavily distorted" |
| Distortion | n | 66600093 | 0 | 0 | distortion#7 | effects applied to electrical instruments to modify their sound output, usually by increasing their gain, "there's so much distortion in their sound" |
| Brutal | a | 66600094 | 0.25 | 0 | brutal#5 | music that is especially heavy or violent, "good death metal is so brutal" |
| Clean | a | 66600095 | 0 | 0 | clean#19 | a description of some element of music, often the vocals, that is without distortion and not overly rough or harsh, "the vocals are clean in a lot of power metal" |
| Dark | a | 66600096 | 0.25 | 0 | dark#12 | a description of music, often heavy forms, that is aggressive or sad in aspect, "this music is so dark" |
| Rubbish | a | 66600097 | 0 | 0.375 | rubbish#2 | something lacking in worth, "this band's music is just rubbish" |
| Dirty | a | 66600098 | 0 | 0 | dirty#13 | a description of music as rough or aggressive, "the guitars sound really dirty" |
| Filthy | a | 66600099 | 0 | 0 | filthy#2 | a description of music as rough or aggressive, "the vocal delivery is absolutely filthy" |
| Comeback | n | 66600100 | 0 | 0 | comeback#3 | a musical release marking the return of an artist after some absence, "they just came out with a new comeback album" |
| Live | a | 66600101 | 0 | 0 | live#3 | music that is played in a live setting, "they made a live album at their last gig" |
| Progressive | a | 66600102 | 0 | 0 | progressive#7 | a descriptor of music that is typically ambitious or complex, "their music is so progressive now" |
| Full-length | n | 66600103 | 0 | 0 | full-length#3 | a full-length music album, as opposed to an EP, single, etc, "their new full-length is due out soon" |
| Thrash | n | 66600104 | 0 | 0 | thrash#2 | a style of music, "they just released a new thrash album" |
| Sludge | n | 66600105 | 0 | 0 | sludge#3 | a style of music, "they just released a new sludge album" |
| Industrial | n | 66600106 | 0 | 0 | industrial#1 | a style of music, "they just released a new industrial album" |

### A.1.2 Modified entries

| Word to change | Synset ID | New scores (pos, neg) |
|---|---|---|
| Ominous | 00194357 | 0.75, 0 |
| Wicked | 02513740 | 0.625, 0.25 |
| Wickedly / Evilly | 00144586 | 0.375, 0 |
| Wickedness | 04827957 | 0.125, 0 / 0.75, 0.125 |
| Wickedness | 04852750 | 0.125, 0 / 0.75, 0.125 |
| Crushing | 00587697 | 0.75, 0 |
| Grind | 01594978 | 0.125, 0 |
| Lethal | 00993885 | 0.125, 0 |
| Lethality | 04791081 | 0.5, 0.25 |
| Evil | 01131043 | 0.875, 0 |

| | | |
|---|---|---|
| Destructive | 00586183 | 0.625, 0 |
| Destructiveness | 05165904 | 0.625, 0 |
| Hateful | 01460421 | 0.667, 0.333 |
| Hatefulness | 04781755 | 0.5, 0.25 |
| Hideous | 00221934 | 0.625, 0.25 |
| Hideousness | 04691061 | 0.375, 0.125 |
| Macabre | 00195684 | 0.625, 0.25 |
| Bitter | 02396098 | 0.375, 0 |
| Bitterness | 07548978 | 0.625, 0.125 |
| Murderous | 00248837 | 0.625, 0 |
| Murderousness | 04830904 | 0.25, 0 |
| Murderously | 00405016 | 0.125, 0 |
| Malevolent | 00225564 | 0.875, 0 |
| Malevolence | 04842313 | 0.875, 0 |
| Bleak / Black | 01229561 | 0.5, 0 |
| Bleakly | 00175255 | 0.25, 0 |
| Bleakness | 14525548 | 0.125, 0 |
| Desolate | 01232507 | 0.75, 0 |
| Angry | 00113818 | 0.5, 0.25 |
| Angrily | 00227323 | 0.125, 0 |
| Forlorn | 01230387 | 0.375, 0 |
| Dirge | 07050619 | 0.125, 0 |
| Elegy | 06379568 | 0.375, 0 |
| Outraged | 00115494 | 0.75, 0 |
| Rage | 13980128 | 0.125, 0 |
| Raging | 01512804 | 0.625, 0.125 |
| Wrathful | 00116245 | 0.25, 0 |
| Damned | 01451225 | 0.375, 0 |
| Damnation | 14458593 | 0.125, 0 |
| Disturbing | 01189386 | 0.875, 0 |
| Creepy | 00195191 | 0.5, 0 |
| Creepiness | 05723080 | 0.125, 0 |
| Disquieting | 00480753 | 0.625, 0 |
| Darkness | 14563564 | 0.75, 0 |
| Manic | 02391003 | 0.625, 0 |
| Maniacal | 02076416 | 0.125, 0 |
| Maniacally | 00395190 | 0.125, 0 |
| Demented | 02075321 | 0.5, 0 |
| Dementedly | 00080890 | 0.5, 0.125 |
| Sombre | 00365261 | 0.625, 0 |
| Dangerous | 02058794 | 0.75, 0 |
| Dangerousness | 04856460 | 0.75, 0.125 |
| Dangerously | 00090228 | 0.375, 0.25 |
| Grim | 01785341 | 0.75, 0 |

| | | |
|---|---|---|
| Sad | 01361863 | 0.75, 0.125 |
| Sadness | 13989051 | 0.625, 0 |
| Gloomy | 00703615 | 0.875, 0 |
| Gloomily | 00232314 | 0.125, 0 |
| Gloom | 14525365 | 0.125, 0 |
| Depressive | 10005163 | 0.625, 0 |
| Doom | 07334206 | 0.75, 0 |
| Cacophony | 07377473 | 0.25, 0 / 0.5, 0 |
| Cacophony | 04984938 | 0.25, 0 / 0.5, 0 |
| Cacophonic / Cacophonous | 00298767 | 0.625, 0 |
| Haunting | 01561079 | 0.375, 0.125 |
| Screaming / Shrieking | 07393161 | 0.125, 0 |
| Wailing | 01365785 | 0.125, 0 |
| Howling | 07126734 | 0.125, 0 |
| Roaring | 07121361 | 0.125, 0 |
| Roaring | 07377682 | 0.125, 0 |
| Furious | 02511528 | 0.375, 0.125 |
| Violent | 02510879 | 0.25, 0 |
| Violence | 13979977 | 0.125, 0 |
| Forbidding | 01802932 | 0.5, 0.25 |
| Blasphemous / Sacrilegious | 02012748 | 0.375, 0 |
| Blaspheme | 00865387 | 0.125, 0 |
| Profane | 02056880 | 0.125, 0 |
| Sacrilegiousness | 04856182 | 0.375, 0.125 |
| Vicious | 00226105 | 0.75, 0 |
| Hellish | 01132515 | 0.875, 0.125 |
| Hellishly | 00132532 | 0.25, 0 |
| Brutalize | 00113853 | 0.75, 0 |
| Brutality / Viciousness | 04830689 | 0.375, 0.25 |
| Monstrous | 00221627 | 0.375, 0.125 |
| Fury | 05037813 | 0.125, 0 |
| Cliché/Cliche | 07154046 | 0, 0.375 |
| Melancholy | 01362684 | 0.625, 0 |
| Terrifying | 00196449 | 0.625, 0 |
| Chilling / Scary | 00194924 | 0.75, 0 |
| Explosive | 01144009 | 0.25, 0 |
| Outrageous | 01626562 | 0.875, 0 |
| Bass | 01215935 | 0, 0 |
| Growl | 07384473 | 0, 0 |
| Growl | 01045719 | 0, 0 |
| Growling | 07210951 | 0, 0 |
| Shout | 07120524 | 0.125, 0 |
| Shout | 00913065 | 0.125, 0 |
| Shout | 00912048 | 0.125, 0 |

| | | |
|---|---|---|
| Wail | 07211950 | 0.125, 0 |
| Wail | 01046932 | 0.125, 0 |
| Roar | 00915605 | 0.125, 0 |
| Roar | 01996188 | 0.125, 0 |
| Howl | 07385367 | 0.125, 0 |
| Howl | 01047381 | 0.125, 0 |
| Scream | 00912833 | 0.125, 0 |
| Scream | 02173336 | 0.125, 0 |
| Flat | 02307563 | 0, 0.25 |
| Genre | 07092158 | 0, 0 |
| Chaotic | 01669507 | 0, 0 |
| Chaotic | 02390569 | 0, 0 |
| Psychedelic | 00086801 | 0, 0 |
| Psychedelic | 01777822 | 0, 0 |
| Seem | 02198234 | 0, 0 |
| Horrifying | 00193480 | 0, 0 |
| Mournful | 01366157 | 0.75, 0 |
| Mournful | 01362387 | 0.75, 0 |
| Raspy | 00299476 | 0, 0 |
| Thundering | 01286375 | 0.25, 0.125 / 0.125, 0 |
| Thundering | 01922030 | 0.25, 0.125 / 0.125, 0 |
| As | 00022131 | 0, 0 |
| Be | 02604760 | 0.125, 0.125 |
| Simply | 00004967 | 0, 0 |
| Serious | 00748359 | 0, 0 |
| Previous | 01729819 | 0, 0 |
| Specifically | 00041758 | 0, 0 |
| Last | 00349894 | 0, 0 |
| Main | 01512527 | 0, 0 |
| Tom | 09638245 | 0, 0 |
| Generally | 00041954 | 0, 0 |
| Single | 00539389 | 0, 0 |
| Already | 00031798 | 0, 0.125 |
| Chord | 06869951 | 0, 0 |
| Rip | 01573276 | 0, 0 |
| Nuance | 06606191 | 0.25, 0 |
| Above | 00125993 | 0, 0 |
| Vanessa | 02275921 | 0, 0 |
| Obviously | 00039318 | 0, 0 |
| Dull | 00391672 | 0, 0.25 |
| Forgettable | 01040239 | 0, 0.5 |
| Amaze | 00724832 | 0.25, 0 |
| Departure | 07366289 | 0, 0 |
| Might | 05030680 | 0.375, 0 |

| | | |
|---|---|---|
| Various | 02065665 | 0, 0 |
| Various | 02067491 | 0, 0 |
| Versatile | 02507772 | 0.25, 0 |
| Sure | 02302822 | 0.125, 0 |
| Death | 13962498 | 0, 0 |
| Death | 15143477 | 0, 0 |
| Death | 09488259 | 0, 0 |
| Adult | 01321456 | 0, 0 |
| Happen | 00344174 | 0, 0 |
| Indecision | 04866866 | 0, 0.25 |
| Indecision | 05699172 | 0, 0.25 |
| However | 00028424 | 0, 0 |
| Magic | 01576071 | 0, 0 |
| Speech | 07071483 | 0, 0 |
| Horror | 03537866 | 0, 0 |
| Monster | 10109443 | 0, 0 |
| Vibe | 04727883 | 0, 0 |
| Black | 04960277 | 0, 0 |
| Black | 13983807 | 0, 0 |
| Black | 00392812 | 0, 0 |
| Black | 00114797 | 0, 0 |
| Black | 01131935 | 0, 0 |
| Black | 00274068 | 0, 0 |
| Black | 02079507 | 0, 0 |
| Originality | 04800359 | 0.375, 0 |
| Extreme | 01511520 | 0, 0 |
| Extreme | 01534858 | 0, 0 |
| Inanity | 05174023 | 0, 0.5 |
| Prophetic | 01881696 | 0, 0 |
| Mediocrity | 04795252 | 0, 0.375 |
| Gothic | 00969103 | 0, 0 |
| Hell | 05629682 | 0, 0 |
| Hell | 01260731 | 0, 0 |
| Hell | 00736786 | 0, 0 |
| Christian | 00411009 | 0, 0 |
| Lucifer | 09543353 | 0, 0 |