Interim Report

# Chapter 1 - Introduction

To determine whether there is a link between the sentiment of a TV show and the viewer ratings. This project will explore areas of Sentiment Analysis (SA), different ways to perform sentiment analysis, why is this relevant and how will others benefit from this. This will involve some data analysis and manipulation to find out if there are any correlations. This chapter explores the background of the research as well as a justification for it. The aims and objectives are also considered.

## 1.1 - Background

Sentiment analysis has been defined as opinion mining (Ding, et al. 2008) and according to Feldman (2013), sentiment analysis is used to look at the “decision-making process of people”. The value of this is to allow users to better understand people as consumers, voters, reviewers etc.

Feldman (2013) states that by using sentiment analysis it “offers these organizations the ability to monitor the different social media sites in real time and act accordingly”. This would give companies a much better understanding of their customers and can benefit from this.

## 1.2 – Aims & Objectives

The aim of this project is to explore the areas of sentiment analysis and to create a script which will look at the sentiment of an episode of a TV show and the viewer ratings and see if there is a link between them.

* To perform a literature review of sentiment analysis.
* To investigate the sentiment of a TV show, per episode/season.
* To investigate the viewer rating of a TV show from reviewer websites.
* To investigate if there is a link between both results.

By researching other methods of sentiment analysis, this would enable this project to try and cover all the different parts of sentiment analysis using each method to their advantages.

# Chapter 2 – Lit Review

According to Pang & Lee (2008) sentiment analysis has also been referring to it as ‘brand monitoring,’ ‘buzz monitoring’ and ‘online anthropology,’ to ‘market influence analytics,’ ‘conversation mining’ and ‘online consumer intelligence’.

## 2.1 - What does it do

Sentiment analysis is a method of analysis which looks at the emotion of a word with the positivity and negativity of the said word. This style of analysis is used in marketing to measure the reviews of a service or product with the product reviews which is also what Taboada, et al (2011) states.

## 2.2 - Types of sentiment analysis

There are multiple types of sentiment analysis, which looks at different types of entities within a data set. These different types are called: Document-level sentiment analysis, Sentence-level sentiment analysis, Aspect-based sentiment analysis, Comparative sentiment analysis and Sentiment lexicon acquisition Feldman (2013).

### 2.2.1 – Document-level sentiment analysis

The first type of sentiment analysis which will be explored is Document-level. This type of sentiment analysis is known as the simplest form of as it looks at the whole document as one attribute (Feldman 2013). For an example of this, looking at different types of reviews from Amazon and would give an overall rating. This type can also be done with machine learning which consists of supervised and unsupervised learning (Bibi 2017).

Supervised sentiment analysis considers such algorithms As Bibi (2017) pointed out, “Naive Bayes, Maximum Entropy classification and Support Vector Machines (SVM).”

* Naive Bayes – has real time prediction, is very fast algorithm.
* Maximum Entropy Classification
* Support Vector Machines

With unsupervised, this approach is a little bit different. As it would need to have been given a certain threshold for the semantic orientation (SO), this would be a level of positivity to make is overall positive or under making overall negative.

Advantages

* Can easily look at a document and give it an overall sentiment score.
* Can be done quickly.

Disadvantages

* Difficult to learn the supervised methods.

Conclusion

Whilst this type of sentiment analysis can be complicated to learn, the unsupervised method could prove useful for the rest of this project. Especially if this was done for each episode of a tv show to get an overall sentiment rating, episode 1 could have a chart showing the top 5 sentiments and with this data, it can be compared to the top 5 sentiment scores from the reviewer datasets.

### 2.2.2 - Sentence-level sentiment analysis

The second type of sentiment analysis is Sentence-level. This type looks at each sentence as an individual entity, so will break down each sentence into an ‘opinion’. Looking at the emotion of each sentence and will show the overall sentiment at the end and how much the sentiment can differ between sentences, from positive, negative or neutral. This type of sentiment analysis would usually be used for the subjectivity classification and the sentiment classification (Bibi 2017).

Advantages

* Good if comparing multiple sentences from the same person.
* Good to see how a person’s sentiment can change over time in their writing/review.

Disadvantages

* Some reviews could be a lot more than once sentence.

Conclusion

Whilst this type of sentiment analysis can be in theory be done for this project, It would depend on the datasets format. Whilst it could lay each review out as its own sentence, some reviews could be much longer than a sentence.

### 2.2.3 - Aspect-based sentiment analysis

Aspect-based sentiment analysis is also known as feature-based sentiment which as stated by Feldman (2013) is used to identify the sentiment of many attributes. Which can be useful when a person is talking about an overall experience but has different experiences at different parts. For example, when looking at a review of a tv episode, the reviewer could have liked the first section of the show, disliked the middle part of the show and really enjoyed the last section of the show. With this type of analysis, the data scientist can pin point the sentiment for each section of the show, to track how the sentiment is changing.

Advantages

* Can be used to see how the sentiment score changes over time.

Disadvantages

* Might become complex when dealing with mass datasets
* Can be difficult to find the features relevant to the individuals work as stated by

Conclusion

This type of analysis could prove to be useful for this project by analysing the reviews sentiment over time and seeing if it follows the sentiment of the tv show over time. This would have to depend on the reviewer’s feedback style and seeing if they touched on different parts of the show or just gave an overall review.

### 2.2.4 - Comparative sentiment analysis;

Comparative sentiment analysis looks at the sentences which are comparing a product/service to a similar product/service. This is used due to the number of reviewers who often compare x to y, here is an example found on amazon.co.uk. For example, “The Samsung J5 has more to offer at half the price. In my view.” – (Google Pixel 3 Review – Amazon.co.uk (2019).

This would be great for comparing 2 different episodes of a tv show. and seeing how they compare to one another, as stated by Feldman (2013) looks to extract the comparative entities which such as better, faster, lighter, sadder in each opinion.

Advantages

* Can be used to compare 2 different entities.

Disadvantages

* Can depend on the dataset formatting.
* May need additional pre-processing.

Conclusion

This type of analysis would be great for comparing two different episodes of a show for this project but would need to be explored with sample sets. This again would depend on the reviewer’s formatting for their reviews, as some reviews might not have any comparative entities to be extracted.

### 2.2.5 - Sentiment lexicon acquisition

Lexicon based sentiment analysis is the most crucial resource (Feldman 2013), this is due to the use of dictionaries which can be hand coded and unique for a specific use case. Alternatively, the dictionaries can be crowd sourced, such as Bag of Words which uses a dictionary of positive and negative words which are all matched up against a score. This is done by following a calculation of:

$$\sum{positive\_matches} - \sum{negative\_matches}$$

Which was pointed out by LyonEye (2016), the scores are then normalised to the form of 1 to 5. There are also other dictionaries the user can choose from such as WordNet, which is described as a ‘Large lexical database of English nouns, verbs, and objectives’ (WordNet 2019).

Whilst being such a crucial resource for sentiment analysis, it requires some calculation of words for a document. Whilst you could use a dictionary if it doesn’t contain the required words for the specific use case, some additional steps may be required. Like creating a custom dictionary to remove stop words and add additional words.

Advantages

* The use of crowd source and premium dictionaries makes sentiment analysis faster.

Disadvantages

* Would need to add a ‘custom dictionary’ for any works which aren’t relevant. i.e. “The”, “I”, “Don’t” etc.
* Would require more training data to get a higher accuracy.

Conclusion

Whilst this method of sentiment analysis is considered the most crucial, it can also prove difficult for when the context starts to get more complex which as stated by Ding, et al. (2018) “This approach allows the system to handle opinion words that are context dependent, which cause major difficulties for existing algorithms”.

### 2.2.6 – Conclusion

Now that the different types of sentiment analysis have been explored, the project can be explored in multiple ways to see what areas would be efficient and relevant to the scope of this project. So far the main methods which stand out are the document level, comparison level and lexicon level. Whilst the aspect level and sentence level have their own use cases, they wouldn’t really be applicable for this type of project.

## 2.3 – Current Software

Sentiment Analysis has become so popular that even the big names in the tech industry have provided their own data analysis tools. In this part of the report current software will be explored, which makes it a lot easier for people to use sentiment analysis and data analysis tools within the work place.

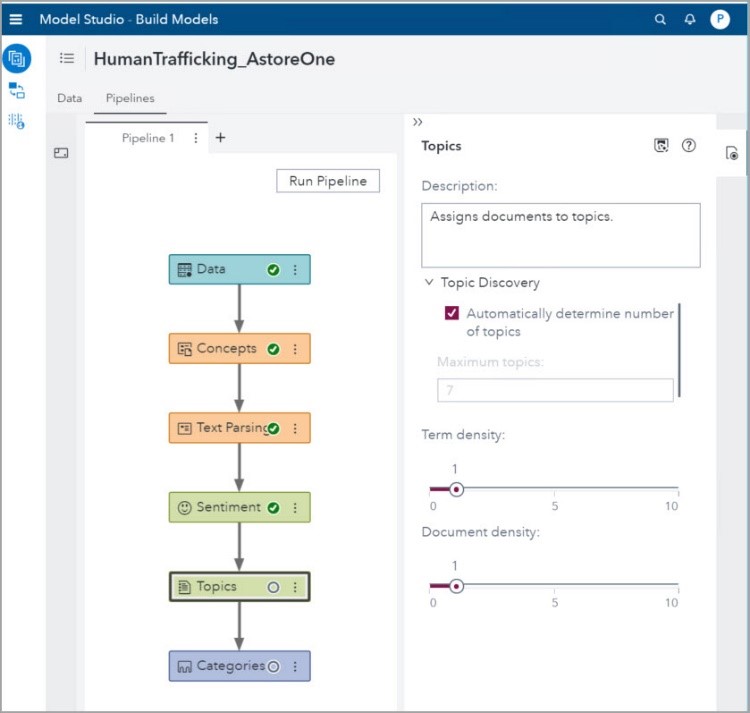
Note: Whilst most of these tools require a premium subscription to use, a lot of them provide a free trial and some of them are provided by the University of South Wales. If there were any premium features used in the report, it would be displayed with a \*.

### 2.3.1 - SAS - Visual Text Analytics - [VTA](https://www.sas.com/en_gb/home.html)

The first sentiment analysis tool this report will explore is Visual Text Analytics (VTA) which was created by a company called SAS, which claims to be the “Analytics Leader” (SAS 2019), by providing tools to execute ‘what-if’ scenarios, sentiment analysis, predictive analytics and confidence intervals (Butler 2015). Whilst a lot of the features provided could be very useful for a business, for the scope of this project this report will focus on the sentiment analysis feature. Their sentiment analysis provides some visuals which will display the sentiment at a document or topic level, whilst offering a neural network to improve accuracy (Neri 2019).

SAS VTA also follows a methodology which was explored in a blog post by Patricia, which can be seen below (in figure 1). This methodology looks like simple steps to follow which could be replicated in an open-source equivalent, using tools like R and R-Studio. From exploring the methodology stated, it isn’t stated that the data has be pre-processed and cleaned which could prove difficult depending on the format of the data.

VTA helps to find the most common topics and terms, for example, talking about themes within tv shows as a topic and the terms would display the most used words from small to large (large being the more popular word). By displaying this information, the user can quickly see their results from all their datasets which has now be processed using VTA.



(Figure 1 - SAS Visual Text Methodology - Neri, P. (2019))

Whilst looking like an easy to use piece of software which would do sentiment analysis easily, by using a drag and drop method, they don’t clearly state a price which could put potential users off. Especially when there are so many other companies doing sentiment analysis, which are as follows:

|  |  |  |
| --- | --- | --- |
| Software: | About: | Conclusion: |
| Watson Tone Analyzer - [WTA](https://www.ibm.com/watson/services/tone-analyzer/) | This software was created by IBM and is used to understand emotions, tone and communication. By doing this, you’re able to get the overall tone (sentiment) from the document, before and after it is published, to make sure the desired emotion the user wants to get across, gets across to their audience correctly.  WTA can also be used to be integrated with companies Chat Bots, which are becoming more popular and out performing humans in the sales industry (Dermoudy 2018). By using WTA with your chat bot, companies can get a better understanding of what type of language and writing style could be used for their customers. | This type of software seems like it was built for scalability, with the inclusion for using it with their chat bots is very interesting. Even if this would mainly be beneficial for business, it still offers the great tools needed for sentiment analysis.  From using their demo, it looks easy to use and the results are displayed at Document-level and Sentence-level. This software could be useful to test against the open-source methods. |
| Google Cloud Natural Language - [GCNL](https://cloud.google.com/natural-language/) | Googles take on sentiment analysis is called Google Cloud Natural Language (GCNL), uses a machine learning approach to “reveal the structure and meaning of text” (Google Cloud 2019). Uses an automated natural language process, which enables the user to train the base model to fit into your work place easily before the evaluation process.  GCNL analysis the sentiment at document-level, sentence-level & entity-level which allows the user to get multiple sentiments out of a dataset.  GCNL offers pre-trained models which can be used without the need of training datasets, which would allow the user to analyse their data at a high accuracy right as they start. | GCNL has a lot to offer with their Natural Language, with 2 different types of API’s for different use cases. With their high accuracy with pre-trained model allows the user to get on with their work, knowing they will have a high accuracy.  Google have also stated their price in an easy to read table, which isn’t displayed on the other options. |
| Amazon Comprehend - [AC](https://aws.amazon.com/comprehend/) | Amazon Comprehend (AC) offers a natural language process service which like the others, offers a machine learning method to quickly and easily analyse datasets. AC claims to take any unstructured datasets and can analyse their sentiments to get “better answers from your text” (Amazon 2019).  Their sentiment analysis offers back the sentiment and the score, with a higher number being better ([Appendix 1](#_Appendix_1)). Whilst they don’t offer the type of sentiment analysis they offer, it looks to follow a similar method of Google from first look but isn’t declared anywhere. | AC offers an easy to use API like GCNL which also offers the price on their website. The price ranges are varying from $1(£0.77p) per job.  It looks like this type of sentiment analysis must be used on the Amazon Web Services, which is pushing their customers onto their platform.  Whilst sounding like a good option, this area should be explored more before deciding on whether to compare it to the open-source route, depending on the scope of the project. |

# 3 – Identify Outputs

From the literature review, there have been many discoveries about sentiment analysis. These are as follows;

Sentiment analysis is very popular for business who want to analyse their customers data, to discovery if there are any patterns which can be found.

Sentiment analysis can be used in multiple ways from Document-level sentiment analysis, Sentence-level sentiment analysis, Aspect-based sentiment analysis, Comparative sentiment analysis and Sentiment lexicon acquisition.

There are already a few companies offering a commercial version which does sentiment analysis. (SAS, SPSS, Google Cloud Natural Language, Watson Tone Analyzer & Amazon Comprehend as well as others).

How this can be implemented with open source tools such as R & R-Studio. with some packages like BagOfWords and Tidytext to do sentiment analysis yourself with a script. Whilst being a free option, it also leaves more room for user error.

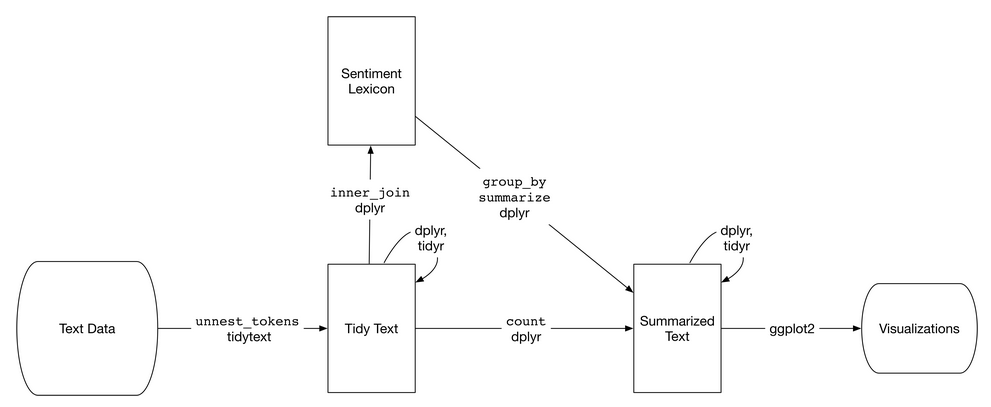
## 3.1 - How they could be applied to the design of the final deliverable

From discovering these different types of sentiment analysis, there is a low barrier of entry for someone who is willing to learn the free and open source approach as declared above. By doing so allows the user to explore the different ways of doing sentiment analysis by coding it themselves and seeing the differences. Whilst doing the user can test different types of sentiment analysis on different types of datasets and see the differences.

This will allow this project to be fully explored down the free route, especially for the scope of this project.

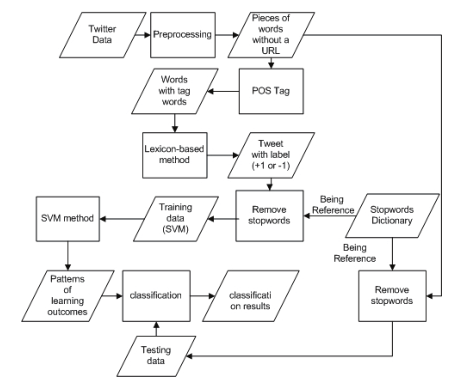
## 3.2 - Methodologies

When researching different types of methodologies and different ways to do sentiment analysis, there have been different ways users have broken down the words and then analysed them. For example, if we look at Silge’s book on the package tidytext (Silge, J. & Robinson, D. 2017) she put the following:



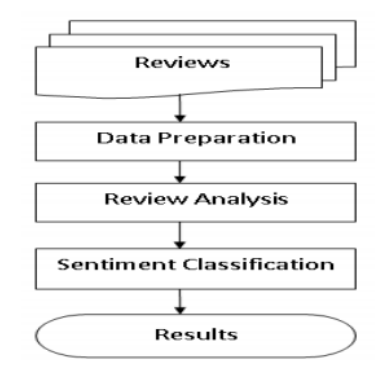
(Figure 2 – Flowchart of sentiment analysis with tidytext, Silge, J. & Robinson, D. 2017)

With these steps, you can clearly see what must be done at each step of the process of cleaning the data and analysing it. Whilst being an easy flow chart to follow, if you were to look at a bigger project such as that of Tiara *et al.* (2015), We can see that from figure 2, the flowchart starts to become a bit more complex.



(Figure 3 – Flowchart of sentiment analysis of twitter data, Tiara *et al.* 2015)

From both diagrams shown, this can be compact into a simple set of rules to follow whilst attempted to perform sentiment analysis. It would be as follows:



(Figure 4 – Sentiment Analysis Model, Bibi (2017))

# 4 - LSEPI

## 4.1 - Legal

Software licenses – The aim of this project is to use open source tools which will give anyone the ability to follow along with this project. Whilst exploring the additional software already available, the licences would be required by the user whom is carrying out the research.

R - A programming language for statistical computing which supports graphics for displaying your data and results.

R-Studio – An open source front end for the programming language R, which is great for creating and manipulating scripts and data frames.

Git - is an open source version-control system for keeping track of changes in code.

GitHub is a website for developers to upload their code externally, which was built on Git and allows for collaboration. (Microsoft)

Visual Studio Code - An open source text editor for developers.

## 4.2 - Social

Anonymity of user input – With the use of web scrapping for this project, this could be a difficult issue for some websites. For the types of websites this project will focus on, most of them offer a developer version which allows users to download and analyse their datasets. This has been confirmed in the Terms Of Service (TOS) and usually requires the user to create an account and tell the company why/what you’ll be doing with the data. This has been done for this project and by doing so with a developer account, allows the user to get a certain amount of data per day.

By anonymising the user’s data, this project will be able to easily share the findings without worrying about the privacy concerns and GDPR. This can be easily done by following what was stated by UK Data Service (2019), applying pseudonyms, generalising information (such as location) and blurring image or video data.

## 4.3 - Ethical

To lay out the rules for ethics, we would have to consider how personal it can be from person to person. Whilst it can be so unique, usually the work place would follow some general ethical concerns.

* To treat people fairly
* To respect the autonomy of individuals
* To act with integrity
* To seek the best results

This project will make sure to be mindful of ethical issues, for example with scraping data it could be easy to identify someone if their tweet wasn’t scrambled up. By doing so, this could cause some back lash (Witch hunt) for someone’s opinion online and could lead to a much bigger ethical issue.

This project will remove all bias from subjects talked about and will be ethical to any information from users and websites. Whilst also following the ethics which has be stated by the University Of South Wales, which are as follows:

* Treat people fairly
* Respect the autonomy of individuals
* To act with integrity
* Seek the best results

It is also important to be mindful of morals when it comes to ethics, as ethically it could be correct to do something, but morally it could be on the fence. This is because morals are even more personal than ethics, as morally someone could do something which would hard others but benefit themselves massively.

## 4.4 - Professional

As stated by the BCS code of conduct, a professional should:

* Only undertake to do work or provide a service that is within your professional competence.
* NOT claim any level of competence that you do not possess.
* Develop your professional knowledge, skills and competence on a continuing basis, maintaining awareness of technological developments, procedures, and standards that are relevant to your field.
* Ensure that you have the knowledge and understanding of Legislation\* and that you comply with such Legislation, in carrying out your professional responsibilities.
* Respect and value alternative viewpoints and, seek, accept and offer honest criticisms of work.
* Avoid injuring others, their property, reputation, or employment by false or malicious or negligent action or inaction.
* Reject and will not make any offer of bribery or unethical inducement.

(BCS 2019)

# References

Amazon (2019) ‘Amazon Comprehend’ Available at: <https://aws.amazon.com/comprehend/> (Accessed 22/11/19)

Amazon.co.uk. ‘Google Pixel 3’ Available at: <https://www.amazon.co.uk/Google-Pixel-SIM-Free-Smartphone-Black/dp/B07K3477FP/ref=sr_1_1?keywords=pixel+3&qid=1574241946&sr=8-1#customerReviews> (Accessed 20/11/19)

BCS (2019) ‘BCS, The Chartered Institute For It Code Of Conduct For Bcs Members’. Available at: https://cdn.bcs.org/bcs-org-media/2211/bcs-code-of-conduct.pdf (Accessed 18/11/19)

Bibi, M. (2017) ‘Sentiment Analysis at Document Level’. Available at: <https://www.researchgate.net/publication/320729882_Sentiment_Analysis_at_Document_Level> (Accessed 30/10/2019)

Butler (2015) ‘SAS Visual Analytics Review’. Available at: <https://www.butleranalytics.com/sas-visual-analytics-review/> (Accessed 21/11/19)

Dermoudy, A. (2018) ‘How AI is Changing the Face of Customer Service’ Available at: <https://www.entrepreneur.com/article/321730> (Accessed 22/11/19)

Ding, X., Liu, B., & Yu, P, S. (2008) ‘A Holistic Lexicon-Based Approach to Opinion Mining’. Proceedings of the 2008 International Conference on Web Search and Data Mining., Pages 231-240, Palo Alto, California, USA. February 11 - 12, 2008

Feldman, R. (2013) ‘Techniques and Applications for Sentiment Analysis’. Communications of the ACM, vol. 56, no. 4

Google Cloud (2019) ‘Natural Language’ Available at: <https://cloud.google.com/natural-language/> (Accessed 22/11/19)

LyonEye (2016) ‘Lexicon-based Bag of Words Sentiment Analysis’. Available at: https://smartcity.readthedocs.io/en/latest/BOW/ (Accessed 17/11/19)

Pang, B. & Lee, L. (2008) ‘Opinion mining and sentiment analysis’. Foundations and Trends in Information Retrieval, 2(1-2), pp. 1-135

SAS (2019)‘Analytics Software & Solutions’. Available at: https://www.sas.com/en\_gb/home.html (Accessed 17/11/19)

Silge, J. & Robinson, D. (2017) Text Mining with R. 1st ed. O'Reilly Media

Silge, J. (2019) ‘tidytext’. Available at: https://github.com/juliasilge/tidytext (Accessed 19/11/19)

Taboada, M., J. Brooke, M. Tofiloski, K. Voll and M. Stede. (2011) ‘Lexicon-Based Methods for Sentiment Analysis’. Computational Linguistics, 37(2), pp. 267-307

Tiara, M. K. Sabariah and V. Effendy, "Sentiment analysis on Twitter using the combination of lexicon-based and support vector machine for assessing the performance of a television program," 2015 3rd International Conference on Information and Communication Technology (ICoICT), Nusa Dua, 2015, pp. 386-390

UK Data Service (2019) ‘Anonymisation’ Available at <https://www.ukdataservice.ac.uk/manage-data/legal-ethical/anonymisation/qualitative.aspx> (Accessed 21/11/19)

# Appendices

## Appendix 1



(Amazon Comprehend – Sentiment Analysis – Amazon (2019))