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 that we can 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, we could look at different types of reviews from Amazon and would give you 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 we could compare it to the top 5 sentiment scores from the reviewer’s 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 data sets 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

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.

“It has the best camera that I have ever used on a phone, keeping in mind that previously I had the Samsung S9.” – (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. With this type of analysis, the user can see if there are any comparative text such as ‘better, happier, sadder etc.’ and see the sentiment scores. When looking at comparative analysis, there is a very interesting paper which helped to improve the accuracy of comparative analysis

Comparing a product to another & Looks for words like, (More, less, lighter) – Look for an existing example of comparative analysis!

FIND REAL EXAMPLES

Advantages

Disadvantages

Conclusion

### 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).

FIND REAL EXAMPLES

Advantages

Disadvantages

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 we will explore these different types of software, 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. Which If I use any premium features to show in this report, will be displayed with a \*.

The first sentiment analysis tool this report will explore is SAS which claimed to be the “Analytics Leader” (SAS 2019).

SAS - VISUAL TEXT ANALYTICS

https://www.sas.com/en\_us/software/visual-text-analytics.html

SPSS – Sentiment Analysis

https://www.ibm.com/analytics/spss-statistics-software

Google Cloud Natural Language

https://cloud.google.com/natural-language/

Watson Tone Analyzer

https://www.ibm.com/watson/services/tone-analyzer/

Amazon Comprehend

https://aws.amazon.com/comprehend/

# 3 – Identify Outputs

**clearly identify the outputs produced from the research/literature review process**

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.

**How they could be applied to the design of the final deliverable**

From discovering these different types of sentiment analysis, there is an easy barrier of entry for someone who is willing to learn the 4th note as declared above. By doing so we can explore the different ways of doing sentiment analysis with coding it ourselves and seeing the differences. Whilst doing this we can test different types of sentiment analysis on different types of datasets.

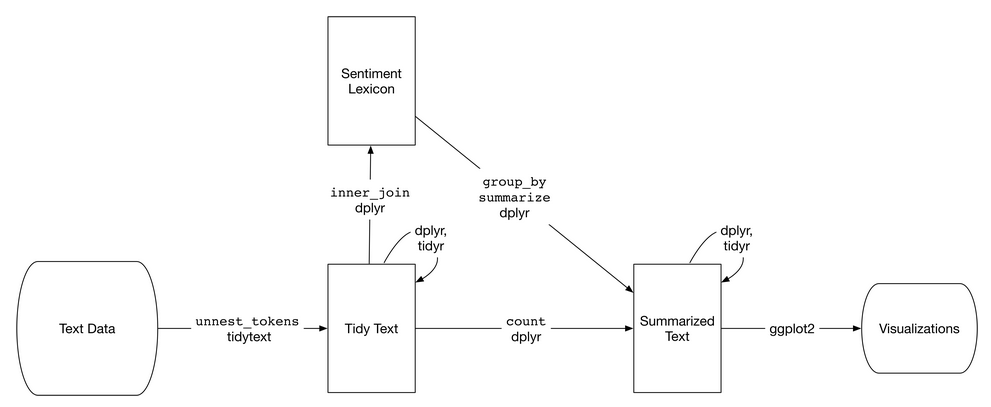
This will allow us to fully explore the free route, especially for the scope of this project.

**Evaluating Software**

**Platforms**

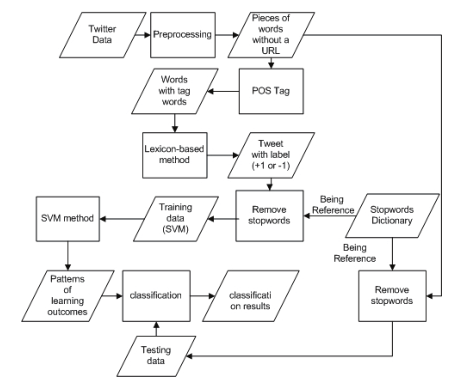
**Methodologies**

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:



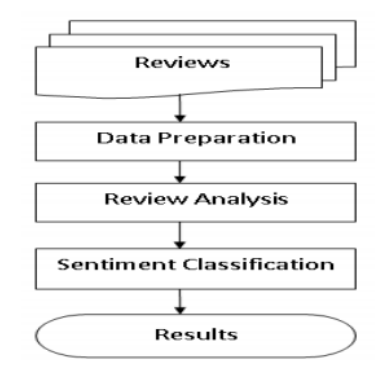
(Figure 1 – Flowchart of sentiment analysis with tidytext, Silge, J. & Robinson, D. 2017)

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 2 – 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 3 – Sentiment Analysis Model, Bibi (2017))

Research will include aspects of:

**evaluating software**

**platforms**

**methodologies**

**academic research**

**evaluation of commercial alternatives**

A chapter should also be included that details the LSEPI aspects associated with your project.

# LSEPI

## 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.

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

R-Studio -

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.

## 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 data sets. 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.

## 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 say 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.

## 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)

## Issues

# References

Behdenna S., Barigou F., Belalem G. (2016) Sentiment Analysis at Document Level. In: Unal A., Nayak M., Mishra D., Singh D., Joshi A. (eds) Smart Trends in Information Technology and Computer Communications. SmartCom 2016. Communications in Computer and Information Science, vol 628. Springer, Singapore.

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)

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.

Jindal, N. & Liu, B. (2008) ‘Opinion Spam and Analysis’. Available at: https://www.researchgate.net/publication/200044297\_Opinion\_Spam\_and\_Analysis (Accessed 16/11/19)

Lexalytics (2019) ‘Sentiment Analysis Explained’ Available at: https://www.lexalytics.com/technology/sentiment-analysis (Accessed 18/11/19)

Lima, A, C, E, S., & de Castro, L, N. (2012) ‘Automatic Sentiment Analysis of Twitter Messages’. Fourth International Conference on Computational Aspects of Social Networks (CASoN). November 2012.

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

Nair, V. (2017) ‘The rise of big data’. Available at: https://www.bcs.org/content-hub/the-rise-of-big-data/ (Accessed 28/10/2019)

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. ‘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.

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)

http://ataspinar.com/2016/01/21/sentiment-analysis-with-bag-of-words/

https://towardsdatascience.com/supervised-vs-unsupervised-learning-14f68e32ea8d