University of Dublin

Trinity College

Sentiment analysis of Irish News Media

Ronan Mac Fhlannchadha
B.A. (Mod.) Computer Science and Business
Final Year Project April 2019
Supervisor: Prof. Khurshid Ahmad

School of Computer Science and Statistics
O'Reilly Institute, Trinity College, Dublin 2, Ireland

Declaration of Authorship

I, Rónán Mac Fhlannchadha, declare that the following report, except where otherwise stated, is entirely my own work; that it has not previously been submitted as an exercise for a degree, either in Trinity College Dublin, or in any other University; and that the library may lend or copy it or any part thereof on request.

Signed:

Date:

Abstract

Public relations is a thriving industry today as organisations recognise the importance of curating and maintaining healthy relationships with the public. Coordinating with the media is an essential part of public relations that helps frame the organisation's image in the mind of the consumer. However, this effect can occur at any time the news media discusses an organisation, not just when efforts are coordinated. This thesis looks at one organisation in particular, An Garda Síochána, and sets out to examine which variables have an impact on the language the news media uses when writing articles concerning An Garda Síochána.

To investigate this relationship, a corpus consisting of 41,779 articles related to the Gardaí was collected from 23 different Irish newspaper from the period January 2017 to December 2018. Six specialist dictionaries were created to perform organisation specific analysis; these dictionaries were: Garda People, Garda Stations, Garda Departments, Crime, Courts and Judges. Through the use of natural language processing (NLP) techniques, the sentiment of the corpus was analysed through the use of the General Inquirer dictionary and the specialist dictionaries.

This data is then used to perform an evidence-based analysis of past behaviour, identifying the degree of sentiment used by Irish news publications and the variables that may be influencing it. This thesis makes use of Vector Autoregression, which provides a statistical means of identifying the degree of causality between one or more variables.

Acknowledgments

I would like to thank a number of people who helped make this thesis the success it is.

First and foremost, I would like to sincerely thank my supervisor Professor Khurshid Ahmad for introducing me to this project idea. Thank you for your continued support throughout the year; your weekly meeting with me provided me with an invaluable source of information. Your wealth of knowledge was evident from day one, and I am lucky and grateful you chose to take me on.

I would like to thank my friends, especially Bobby McGonigle. Your knowledge of R, statistics, and econometrics knows no end and I always took away a lot from our discussions of the Irish news media.

Finally, a massive thank you to my family who supported me throughout the year, especially my father Michael, who proofread my thesis and was a constant source of encouragement. I am the person I am today because of you, without you none of this would be possible.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This chapter will begin with a short introduction of the motivation for carrying out this research project. After this the research question will be proposed, followed with by the research objectives which have been set in order to address this question effectively. This chapter will conclude with a summary of the structure of the remainder of the thesis.

1.1 Motivation

Reputation management is a primary focus for many organisations today, especially public bodies such as An Garda Síochána, in the judgment of their operational effectiveness. In the Garda annual report, improving public opinion regarding the ability of An Garda Síochána to tackle crime was a key performance indicator which they are putting considerable focus on. The target set was to increase this KPI from 57% in 2015 to 60% in 2017, however, as the Garda annual report shows, this number decreased to 55% (An Garda Síochána 2017). In an attempt to have an accurate understanding of the public perception and what may be causing it to fluctuate, An Garda Síochána commission Amarách Research to carry out a quarterly and annual survey continually. These surveys are conducted by means of a face-to-face interview across 200 sampling points.

However, the surveys are limited by the manpower which can be invested in each survey, which currently stands at 1,500 respondents per quarter. Furthermore, face-to-face surveys only give you opinions at a particular time, as An Garda Síochána conduct quarterly reviews, this is only four times a year. It does not show how opinions may fluctuate in between these times. For example, we see from the Q4 2018 survey that 66% of respondents believe

that the Gardaí are effective in tackling crime, it is possible that this opinion may be heavily influenced due to significant crime reporting which occurred the day prior (An Garda Síochána 2018*b*). These surveys are also susceptible to bias, such as social desirability bias due to the presence of an interviewer as identified by Duffy et al. (2005).

Similarly to the analysis of a company from a financial perspective, one's perception of the performance of An Garda Síochána can be done on a fundamental and technical level. Fundamental analysis looks at the physical characteristics of the organisation, such as the physical visibility of Garda personnel, vehicles and stations. Alternatively, technical analysis looks at statistics and trends of the Gardaí which can be seen in the monthly report to the policing authority as required under Section 41A of the Garda Síochána Act 2005. However, what both of these methods of analysis fail to recognise is that the ultimate judge of performance is done by human-beings. This is the central focus of the rapidly rising area of behavioural finance which investigates the impact our emotions and psychology have on our judgement. This judgement of ours is seldom fixed and can often fluctuate from day-to-day as our experiences shape our opinions.

The media and the digital world provides vast amounts of untouched data which can be mined to extract valuable insights and uncover hidden relationships between unexpected variables. Social media sites such as Facebook and Twitter can be a source of consumer opinion in which the people can truly express their opinion with limited influence from external forces. Seeing this availability of massive, raw, and unfiltered data as well as the growing body of behavioural finance research, this thesis sets out to attempt to automate the process of performance analysis for performance analysis.

1.2 Research Question

The precise research in questions for this thesis can be summarised as follows:

To what extent can the sentiment about an organisation in the news be related to external variables?

1.3 Research Objectives

To effectively address the research question several research objectives have been defined:

1. Gather a high quality and substantially large set of documents (corpus) relevant to An Garda Síochána
2. Create an Ontology for sentiment and categories related to An Garda Síochána
3. Perform sentiment analysis on the gathered corpus to identify sentiment and term frequency
4. Analyse the news publications, identifying the type of language which is used and the context in which it is used.
5. Perform Vector Autoregression (VAR) analysis on the results in an attempt to statistically quantify any causal relationships which may exist.

1.4 Report Structure

The next chapter will commence with a review and discussion of the current literature on the effects of the media on the psychology of the public (Chapter 2). The involves the discussion of two effects of the media which are of particular interest for this thesis; the effects of the press on the perception of crime and the effects of the media on the public opinion of an organisations performance. This is followed by a discussion of the current content analysis literature, with a particular focus put on sentiment analysis. The current psychological theories of emotion will be discussed, with a particular emphasis placed on emotion as distinct categories approach and emotion as dimension approach. Furthermore, different approaches to sentiment analysis are discussed, particularly the lexicon-based approach, Machine-learning based approach and hybrid approaches. The chapter concludes with a discussion of statistical methods to be employed throughout the report, ending with an overview of Vector AutoRegression (VAR).

A detailed discussion of the method used will follow, with the method being split into three core area: data acquisition, content analysis and statistical

analysis (Chapter 3). Each sections specific implementation will be discussed in detail. The tools used to perform content analysis, the specialist dictionaries created and the models chosen for VAR analysis will be described and evaluated.

The following section (Chapter 4) provides a detailed discussion of the results obtained from the analysis of the method as described previously. The chapter will commence with an initial review on the corpus which was analysed. Afterwards, the results of the analysis of the corpus through the use of the Harvard General Inquirer (GI) dictionary and six specialist dictionaries related to An Garda Síochána will be discussed. This chapter then concludes with a discussion of the results obtained from the VAR analysis of the model created.

The report concludes with a summary of the method which was carried out for this report (Chapter 5). Subsequently, there will be a brief commentary on the key finding discovered from the previous section (Chapter 4) followed by discussion of some improvements which could be applied to enhance the project, thus setting the stage for future work. The thesis will conclude with a final evaluation, outlining some challenges that were encountered, along with the person achievements gained along the way.

Chapter 2

Literature Review

2.1 Role of the media in influencing perceptions

While it can be difficult to prove, it is widely believed that outside of personal first-hand experiences, what we know, believe, and think about the world around us is largely shaped by how these events are reported to us by the news media (Wilson et al. 1993, Maeroff et al. 1998, Wimmer & Dominick 2013). This concept has been researched by Gerbner et al. (1980) and called 'cultivation theory' which believes that the media has long-term effects on the behaviour of an individual, which are gradual, small, and indirect but cumulative and over time becomes significant in 'cultivating' an individuals conception on social reality. Since its inception, this theory has been heavily researched, with many researchers confirming this theory in their own independent analyses (Mosharafa & Mosharafa 2015). A study in particular by Page Benjamin & Shapiro (1992) over a 15 year period found that television news coverage of major foreign events not only influences the salience of the events but also was a significant predictor into the shift of public opinion towards these events.

An alternative, yet somewhat complementary theory to the cultivation theory is the 'agenda-setting theory'. The theory was first put forward by McCombs & Shaw (1972), but was picked up a few years prior by Bernard Cohen where he states "the press may not be successful much of the time in telling people what to think, but it is stunningly successful in telling its readers what to think about" (Cohen 2015). The theory has since been extensively researched, extended upon, and is still considered relevant today (McCombs & Shaw 1993). The premise of the theory is that the media not only tells us what to think about but through the use of framing as identified by

Gitlin (2003), how to think about it. While this theory was put forward in a political context, it has relevance to this study in the influence of crime perceptions. Lowry et al. (2003) conducted a longitudinal investigation into the application of cultivation theory and agenda-setting theory and identified that news television networked not only made the public concentrate on crime, but was also a key driver in changing the public opinion of crime being the most important problem in the United States from from 5% in 1992 to 52% in August of 1994.

The two theories mentioned above are just two of many which have been created to try and explain the noticeable effects the media has on the public. The following two sections discuss two of these effects in particular: The effects the media has on public crime perception and the effects the media has on public perception of an organisations performance.

2.1.1 Crime Perception

Personal experience of crime is one of the most critical variables a person uses in their judgment of crime levels. However, few experience crime first-hand in their daily lives and therefore must rely on other means to form their judgment. Multiple studies have analysed the media relationship with crime perception and have found it to be a key influencer of public perception (McCullagh 1996, Gerbner et al. 1980, Ramos & Guzmán 2000, Dammert & Malone 2004). O'Connell et al. (1998) investigated whether a causal relationship between the media and public opinion exists in a top-down model (Media shapes public opinion) or bottom-up (Media responds to public opinion). While evidence was found regarding a relationship between them an exclusive top-down model could not be confirmed, a similar finding as identified by Ditton et al. (2004).

As the core function of news media is to report news-worthy topics such as crime, people often see the media as a 'window to the world'. As Gurevitch et al. (1982) states, the news is a socially-manufactured product, which is the result of a selective process from journalists and editors in deciding which stories should be printed and which should not. The factors that each newspaper uses invariably depends on its own selection process, however, as identified by Galtung & Ruge (1965), there are certain 'news-values' which consistently have an effect of increasing the likelihood of a news item becoming a story. One of these news-values, in particular, is "The more negative the event in its consequences, the more probable that it will become a news

item." (Galtung & Ruge 1965). These news items have been modernised lately to include, celebrity and sex, however, violence and crime still remain a major factor (Jewkes & Linnemann 2017).

It is important to note that in the eyes of the media, not all crime is equal. "if it bleeds it leads" - a common adage used within the media industry highlights this point. Roberts (2018) identified that 'news media coverage of crime and punishment reflect a view that is biased towards serious and sensational crime'. Additionally, Williams and Dickinson identified news media coverage of crime is a credible issue to criminologists 'because of the assumption that the salience given to certain types of crime, notably those involving sex or violence, creates a distorted picture of reality which is reflected in the beliefs of news consumers' (Williams & Dickinson 1993). This view is supported by Irish Criminologists, stating how the media tend to "emphasise crimes of violence rather than the more common crimes against property. As a result, levels of fear of crime are higher than they should be" (McCullagh 1996). This has been further shown by Parisi et al. (1979) who found that people are inclined to believe that the crime situation is at least as serious as its portrayal in the media.

Further analysis by Ramos & Guzmán (2000) concluded that while crime rates had risen in the decade in question, the media played a central role in defining fear of crime and specifically fear of certain types of places and even people. This phenomenon is coined 'mean-world syndrome', which states that the medias tendency to provide a biased account of the crime, leads people to adopt an attitude that the world around them is 'mean' and 'dangerous' (Gerbner et al. 1980). A similar theme has been identified by O'Connell (1999) who analysed a sample of 2191 Irish newspapers. O'Connell goes on to describe four specific areas of bias, being "the bias towards extreme and atypical offences in terms of frequency, the bias towards those extreme offences in terms of newspaper space, the bias towards stories involving vulnerable victims and invulnerable offenders and the bias towards pessimistic accounts of the criminal justice system generally" (O'Connell 1999).

It also appears that the degree of trust an individual puts into the press also has an effect on their fear of crime, with a paper by Dammert & Malone (2004) identifying through ordinary least squares regression, that every 1 unit of increase in insecurity led to a 0.44 increase in fear of violence and 0.43 increase in fear of assault and robbery. The previous results were achieved when victimisation and trust in the press were held constant, however, once trust in the press was taken into account, those "who trust the press are 0.26 more fearful of violence, and 0.22 more fearful of robbery or assault, than are

those who do not trust the press (Dammert & Malone 2004). This discovery is further supported by the congruity principle of Osgood & Tannenbaum (1955) who identified the reputation of the source of the news is an important variable used by individuals when forming their own perceptions. This is an important point to note as for many people, newspapers are considered an authoritative source of information, and hence they will have trust in the press. In research conducted by Knowles (1982), 95% of the sample cohort of US residents studied reported that the news media was their most important source of information about crime and numerous additional studies such as Last & Jackson (1988), Graber (1979) have also reached a similar conclusion.

2.1.2 Organizational Reputations

There is currently ample research into the tangible effects of the politico-media complex on political elections. However, there are much fewer empirical real-life studies that focus on the influence of news media on the attitude towards organisations (Carroll & McCombs 2003, Wartick 1992). One such study by Meijer & Kleinnijenhuis (2006) made some statistically significant connections between the news media and organisations reputation identifying that news about the successes of a firm, such as increased performance, improved its reputation. Furthermore it was also discovered that for some organisations, the more a company was criticized by its competitors, the greater its reputation improved. One potential explanation for this is because people want to support the winner, Lazarsfeld et al. (1944) have called this the 'bandwagon effect'.

Meijer & Kleinnijenhuis (2006) studied the relationship between news media and the police in particular and discovered two significant effects that affected the reputation of the police. The first identifies that the bandwagon effect particularly applies to the police, with the more frequent they were in the news with their successes, the greater their reputation improved, and vice versa. The second effect is that support & criticism news has a 'normal' positive impact on the police, where the reputation of the police worsened the more they were criticised in the media, unlike other organisations such as Shell, Albert Heijn, and Schiphol whose reputation increased when criticised. [1].

[1]Support & criticism news is defined by Meijer & Kleinnijenhuis (2006) as occurring when "a company is supported or criticised in the news by another company or by another actor, such as the government or a non-governmental organisation'

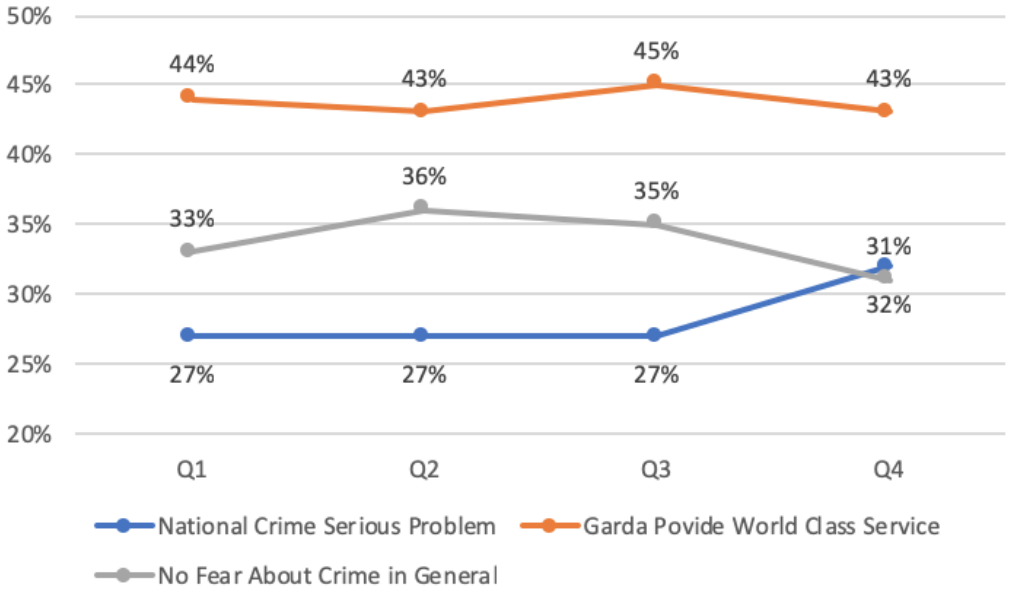

Figure 2.1: 2018 Public Attitudes Survey

The primary function of An Garda Síochána is maintaining law & order; therefore, the perceived crime level is an essential variable as a high perception of crime leads the public to believe there is a breakdown in law & order. As previously discussed, the media has been shown to have a considerable impact on the perception of the levels of crime and therefore has a considerable impact on the reputation of An Garda Síochána. This can be seen in Figure 2.1 which shows an increase in percentage of respondents that believe national crime is a serious problem was followed by a reduction in the percentage of respondents that believe An Garda Síochána is providing a world-class service and a reduction in the percentage that said they have no fear about crime in general.

2.2 Sentiment Analysis

Sentiment Analysis, or opinion mining, is "the field of study that analyses people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organisations, individuals, issues, events, topics, and their attributes." (Liu 2012). Sentiment analysis is a core part of Natural language Processing (NLP) and has been used exten-

sively in major industries such as analysing financial asset returns (Tetlock et al. 2008), analysing political sentiment from social/news media (Ahmad et al. 2011) and gathering business insights from customer reviews and feedback (Turney 2002, Kennedy & Inkpen 2006). (Additional uses see (Liu 2012))

Sentiment analysis can be performed on multiple levels. Document-level is concerned with determining if the overall sentiment of an opinion piece is negative or positive; this is commonly used when analysing reviews (Turney 2002, Pang et al. 2002, Pang & Lee 2004). This is done when it is believed the document only contains one core opinion; however, if it is believed that the document consists of multiple opinions, then the analysis can be done on a sentence or phrase level. Sentence level analysis is very similar to subjectivity classification which distinguishes sentences that contain factual information (objective sentences) from sentences that contain opinionated information (subjective sentences). This level of sentiment analysis is commonly performed when conducting summarisation and multi-perspective opinion analysis (Wilson et al. 2009). The finest grain sentiment analysis is aspect/entity sentiment analysis. Rather than looking at language constructs (paragraphs, sentences, clauses), aspect level looks at the sentiment and the target (opinion). For example "The iPhone's call quality is good, but its battery life is short" analyses two separate aspects, negative aspect for battery and positive aspect for call quality of the target (iPhone) (Liu 2012, Zhang et al. 2011).

Language is inherently littered with ambiguity. People perceive words differently to each other, and it is, therefore, difficult to give a definite value for a words sentiment orientation. For example, when creating a manual corpus, the kappa statistic is often used for determining inter-annotator agreement. Similar to most correlation statistics, the values range from -1 to +1. Under this, a score of 0.61 is considered substantial, and a score of 0.4 is considered moderate and therefore acceptable (Wiebe et al. 2005, McHugh 2012). A further issue with extracting meaning from written text is Zipf's law of meaning distribution. This law states that there is a systematic frequency-rank relationship between the number of meanings a word has and the number of times it appears in text (Piantadosi 2014). Considering that the top 100 words in a frequency-rank collection account for roughly 50% of the total words in a text collection (Ahmad 2008), this implies that any given text will contain a massive amount of potential interpretations. An example of how difficult it is to interpret language can be seen from the sentences below which is often used in linguistics.

Time flies like an arrow. Fruit flies like a banana

To help reduce some of this ambiguity stemming and lemmatising can be used to reduce a term to its inflectional root. Stemming can be achieved by removing the beginning or end of the word resulting in a term which may not have any dictionary meaning (Mogotsi 2010). The most common stemming process is the five-step Porter stemmer however others do exist and can be used as alternatives such as Lovins stemmer and Paice stemmer (Porter 1980). Lemmatising, on the other hand, is the process of reducing a term to its lemma, which can be considered its dictionary form. Both approaches require a predefined dictionary to perform their job. Lemmatisation is generally the preferred way as it ensures the term returned makes sense and can be easily understood post-lemmatisation. However, stemming is a more computationally efficient process and it is not always required to reduce a word to its dictionary term. For example, stemming the term "saw" may result in the character "s" being returned, while lemmatising the term saw may reduce "see" or "saw". These approaches can be used along with Part-of-speech (POS) tagging and word sense disambiguation. Part-of-speech tagging tags each term in a corpus accordingly to their speech tags, e.g. noun, pronoun, adverb, adjective, verb.

Part-of-speech tagging has been shown to be an effective means of disambiguating terms, with Stevenson and Wilks achieving 94.7% accuracy in 2001 (Navigli 2009). The UPenn TreeBank II is a commonly used tag set and contains 36 POS tags and 12 other tags for punctuation/currency symbols.[2] Rule-based taggers such as that by Brill (1992) were the first to appear, however, since then stochastic POS taggers have been implemented such as those using Hidden Markov Models (HMM). These probabilistic taggers have been shown to produce high accuracy results such as the paper by Brants (2000) which achieved an accuracy of 96.7%.

Table 2.1: Stemming vs Lemmatization

Term	Stemmed	Lemmitized
studies	studi	study
stydying	study	study
having	hav	have
beautiful	beauti	beautiful
beautifully	beauti	beautiful

[2]A Breakdown of the tags can be viewed at www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

Furthermore, there is still no consensus today on how emotion is experienced in humans. The two primary approaches into the cognitive account of emotion are emotions as finite categories and emotion of dimensions (Devitt & Ahmad 2013). The former believes that humans are endowed with a discrete and limited set of emotions, each of which arises due to a unique activation of neural pathways (Ekman 1992). Alternatively, the latter believes that "individuals do not experience, or recognise, emotions as isolated, discrete entities, but that they rather recognise emotions as ambiguous and overlapping experiences." (Posner et al. 2005). This dimensional model can be compared to the spectrum of colours, lacking discrete boundaries which are separating one from the other (Russell & Fehr 1994).

As Russell (1994) highlights, there is continued debate as to which method is preferable, however, at present it is believed the dimensional approach is more accurate as the categorical approach "is not exhaustive and does not cover all emotionally-charged experience or indeed text but rather a subset of discrete non-decomposable emotional states" (Devitt & Ahmad 2013). This is supported by research by Eerola & Vuoskoski (2011) who identified the "poorer resolution of the discrete model in characterising emotionally ambiguous examples". The two primary dimensions which consistently appear in research are the dimension of valence and arousal Watson & Tellegen (1985). Commonly a third dimension, dominance, as proposed by Mehrabian & Russell (1974) is also used in the dimensional model, albeit to a lesser extent (Russell et al. 1981). This report makes particular use of the dimensional model, with a particular focus put on the dimension of valence as it has been shown to have the most significant impact on cognitive processes relative to other emotion dimensions (Niedenthal & Halberstadt 2000).

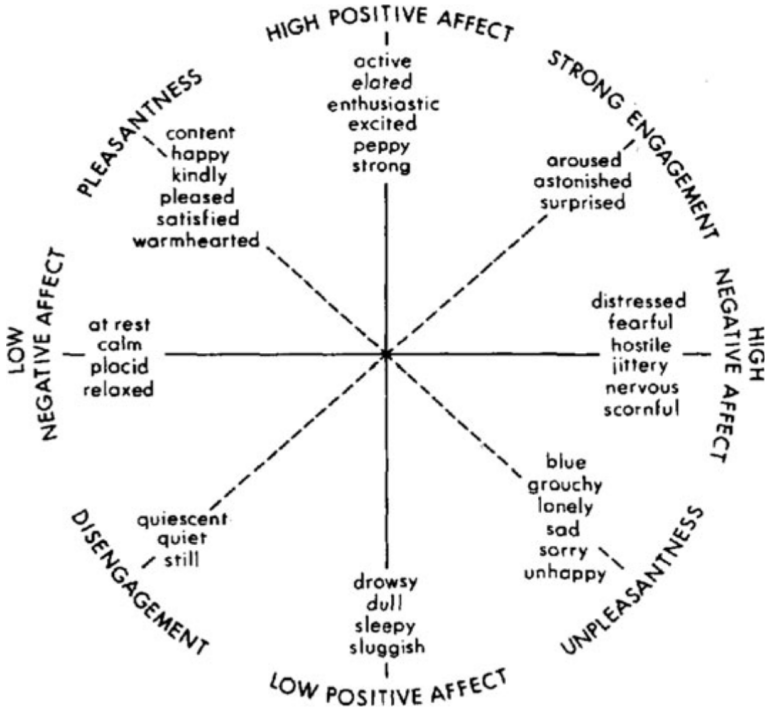

Figure 2.2: The two factor circumplex of affect (Watson & Tellegen 1985)

2.2.1 Approaches to Sentiment Analysis

There are two broad categories of sentiment analysis: Lexicon based and Machine Learning based approaches (Figure 2.3). Lexicon based approached rely on a predefined dictionary with terms annotated with the sentiment polarity and potentially sentiment strength. Machine Learning models, on the other hand, often create classifiers based on predetermined labelled examples called a training set. These methods are not mutually exclusive, and often hybrid approaches are used. The approach chosen depends on the application, domain and language. Lexicon based approaches require a predefined dictionary which may not always be present, especially when operating in a specialist domain. Similarly, Machine Learning models require labelled training data which may not always be possible, while an unsupervised approach may be adopted; this has been consistently shown to give inferior results than supervised or lexicon based approaches (Navigli 2009).

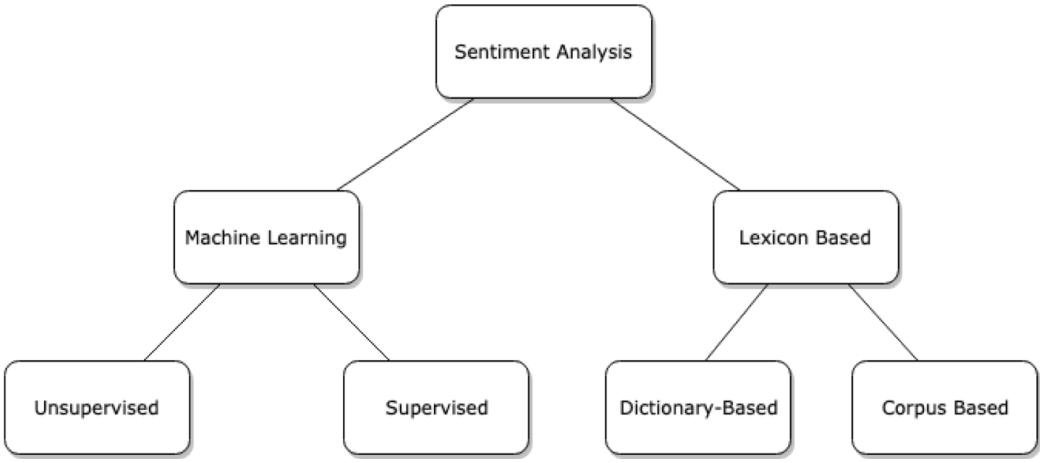

Figure 2.3: Approaches to Sentiment Analysis

Machine-Learning Based

Machine-learning based approaches have risen in popularity since early 2000 (Socher et al. 2013, Kim 2014, Cortes & Vapnik 1995). There at present exists numerous different approaches which can be taken such as Naive Bayes, Support-Vector-Machine, Neural Nets, K-Nearest neighbours, centroid classifier, winnow classifier, and the N-gram model. It is more common to see supervised machine learning for sentiment analysis due to the presence of well-curated manually labelled data. There have been some studies which have shown machine learning methods are superior to lexicon based models for most tasks except topic based classification (Pang et al. 2008). However, in the absence of testing data or when analysing a different domain to the training data, time period or topic, this advantage quickly dissipates (Aue & Gamon 2005, Read 2005).

Perhaps the most simple machine-learning based approach is Naive-Bayes. It is considered somewhat naive as it assumes every word in a sentence is independent of each other, with such an assumption saving massively on computational work. While simple it still proves to be highly effective and still has widespread use (Melville et al. 2009, Xia et al. 2011). It is a probabilistic approach which given a feature vector table, calculates the posterior probability that the sentence/document belongs to each group and assigns it to the group with the highest posterior probability. The posterior probability can be calculated as follows with c_j representing category j and d_i representing document i.

$$p(c_j|d_i) = \frac{p(d_i|c_j)p(c_j)}{d_i} \qquad \text{(Eq. 2.2.1)}$$

Support-Vector Machine is a discriminative classifier which was first introduced by Cortes & Vapnik (1995). It is based on the structural risk minimisation principle which in essence tries to maximise the distance between the two nearest support vectors in each category. While the traditional SVM is applied to linear data on a two-dimensional plane with a binary classification, it is possible to perform SVM on non-linear data and multiple classes through the use of the 'kernel function' or 'one-vs-one/one-vs-all' classification respectively. Multiple forms of SVM exist, although linear SVM has been shown to be extremely performant in text categorisation (Yang et al. 1999).

Neural networks have shown great promise in recent research, with Socher et al. (2013) implementing a Recursive Neural Network (RNN) along with Stanford sentiment treebank which "outperforms all previous methods on several metrics..... pushing the state of the art in single sentence positive/negative classification from 80% up to 85.4% .... and is one of the few models that can successfully handle negation". Furthermore, recent research using a convolutional neural network (CNN) identified that simple CNN with one layer of convolution performs remarkably well against other supervised machine learning techniques (Kim 2014).

Lexicon-Based Approach

Dictionary-based approaches are often praised due to their simplicity and efficiency while still resulting in high levels of accuracy. This approach can also be referred to as the 'Bag-of-words' approach as it splits the document up into token in a process called text vectorisation and discards any semantic relationship between the words, only retaining the word frequencies. It is based on the premise that a documents sentiment orientation is an aggregation of the sentiment orientation of each individual token that the document is comprised of (Yadav et al. 2013). The total polarity of the document is therefore calculated by comparing each word against a predefined polarity dictionary to determine what the overall sentiment orientation is. This approach to sentiment analysis has been used with good accuracy by many researchers, especially in the analysis of the polarity of reviews. (Turney 2002, Pang et al. 2002, Tong 2001)

All studies which utilise this approach require a predefined dictionary with which the tokens can be compared against. These dictionaries are generally made by computational linguistics and are peer-reviewed for quality. Some of which include the General Inquirer (GI) dictionary, which merges the Harvard IV-4 and Lasswell dictionary together (Stone et al. 1966). Additional reputable dictionaries include the Linguistic Inquiry and Word Count (LIWC) developed by Pennebaker et al. (2001) and the opinion lexicon by Hu & Liu (2004). These dictionaries are suitable if working on a general data set which does not contain industry-specific languages, such as that from newspapers who mainly report on facts and lack slang. If working on a specialist industry, such as finance, then a specialist dictionary should be used in addition or instead of the general dictionary, making many of the dictionaries context dependent (Grimmer & Stewart 2013). Loughran & McDonald (2011) identified that nearly three-quarters of the negative words present in the General Inquirer dictionary were not negative in a financial context. In response to this the Loughran-McDonald Master Dictionary was developed, whose importance is seen by Heston & Sinha (2014) who identified that sentiment measures extracted from the GI dictionary and Loughran-McDonald dictionary were in fact negatively correlated.

As mentioned previously, one should always ensure to use a specialist dictionary to ensure the polarity of the words matches the context. In the absence of a specialist dictionary, it can prove costly and time-consuming to construct one for the given context. A possible solution to this is in creating a dictionary from the corpus itself which called the corpus-based approach. There are two corpus-based approaches which commonly cited, the conjunction and co-occurrence approach. Irrespective of which approach is chosen, an initial 'seed-list' must be provided, which is generally a small list (15-25) of universally negative and positive words. It is possible to use a much smaller list, for example, Turney (2002) used only the mutual information between the words 'excellent' and 'poor' and a given phrase for his analysis. The conjunction approach analyses the corpus for words that are explicitly conjoined to words from the seed list and adds them to the dictionary (Rice & Zorn 2013). Once the negative and positive dictionaries are created, they should both be manually scanned to ensure and irrelevant or unsuitable terms are removed. This approach has been effectively used by Hatzivassiloglou & McKeown (1997) having achieved a classification precision of over 90% accuracy on adjectives which appear in a modest number of conjunctions in the corpus. The co-occurrence approach analyses the corpus for all word which co-occurred with words from the seed list. The polarity for each co-occurring word can then be established by deriving the proportion of times it appears with a nega-

tive word, the proportion of times it appears with a positive word and then performing a log odds ratio on the result (Rice & Zorn 2013). Both of these approaches are considered to be minimally supervised as the only require an initial manually created seed-list and a dictionary evaluation at the end.

The dictionaries used in a dictionary-based approach are all called a unigram model as they compare only one word at a time and ignore neighbouring words. The major downside to this approach is that grammar, context, and the positioning of the word in a sentence can affect the polarity of the word. For example, according to a unigram system, 'not bad' and 'very bad' are both equally negative. The words 'not' and 'very' here are considered to be contextual valence shifters (Polanyi & Zaenen 2006). Pang et al. (2002) discovered in the case of movie reviews unigrams performed better, however in spite of this, Dave et al. (2003) identified bigrams and trigrams as yielding better review polarity classification. WordNet, developed by Miller (1995) is a practical approach to circumventing this issue, where "nouns, verbs, adjectives, and adverbs are organised into sets of synonyms, each representing a lexicalised concept" (Miller 1995). In total, WordNet contains 118,000 different word forms and more than 90,000 different word senses. Even so, lexicon based sentiment analysis still has some downfalls such as an inability to handle negation effectively and fails to understand the semantics of the words (Le & Mikolov 2014). Recent research such as Word2Vec by Mikolov et al. (2013) and GloVe by Pennington et al. (2014) use a combination of machine-learning and co-occurrence based sentiment analysis does an effective job at retaining the semantic relationships between words in a corpus. Progress has also been achieved in circumventing the lexicon-based approaches ineffectiveness at handling negation such as algorithms formulated by Turney (2002).

2.2.2 Harvard GI dictionary

The computational speed and accuracy of lexicon-based sentiment analysis along with its intuitive nature makes it an attractive option for performing sentiment analysis in this thesis. The Harvard General Inquirer dictionary, or Harvard dictionary of affect as it is sometimes called, was specifically used in this report to perform sentiment analysis (Stone et al. 1966). There are presently numerous lexical dictionaries of affect to choose from, however, the GI dictionary is often considered to be the 'gold standard' as it a manually created dictionary base on sound psychological experimentation principles. Many researchers have used it to test the effectiveness of their unsupervised

machine learning created dictionary for accuracy (Andreevskaia & Bergler 2008, Taboada et al. 2011). While numerous alternative dictionaries exist as mentioned in the previous section, multiple papers have compared the GI dictionary against other dictionaries of affect, including derivatives of the WordNet, concluding sentiment levels were only marginally affected by choice of dictionary (Devitt & Ahmad 2007, 2008, Ahmad 2011).

The Harvard GI dictionary contains 11,788 rows, with each row containing a one or more assignments to specific categories. The present version is V1.02, which primarily comprised of the Harvard IV dictionary Thorndike-Lorge 1920s–1940s corpus (Thorndike & Lorge 1944) and the Lasswell corpus from pre-1950 but was updated in 1980 (Lasswell 1948, Weber & Namenwirth 1987). It also includes five categories based on the social cognition work of Semin and Fiedler, resulting in 182 categories in total.

A point of criticism for the GI dictionary is that it is a unigram dictionary and therefore does not look at co-occurring words and hence lose context. Taboada et al. (2011) attempted to avoid this issue by creating a new dictionary on top of the GI dictionary called the sentiment orientation calculator (SO-Cal). SO-Cal was developed to be able to handle contextual valence shifters, such as negation and intensification (Polanyi & Zaenen 2006). While they successfully implemented these features Taboada et al. (2011) concluded that the "The General Inquirer lexicon....does comparably quite well despite being relatively small". Further research by Barry (2017) discovered that incorporating a Long Short Terms Memory model along with the bag-of-words model yielded superior results across a range of metrics, however, the standard bag-of-words approach still performed very well. The GI dictionary alone has already been used in various Papers with positive results (Tetlock 2007, Tetlock et al. 2008, Ahmad et al. 2011, Feldman et al. 2010, Kennedy & Inkpen 2006)

Table 2.2: Havard GI Dictionary Categories Used

Category	Num Words
Positive	1,915
Negative	2,291
Strong	1902
Weak	755
Active	2045
Passive	911
Hostile	833
Power	689
Econ@	88
Polit	507

2.3 Statistical Modelling

As the research questions show (Chapter 1), a key focus of this thesis is to examine the language the media uses in its discussion of An Garda Síochána as well as any causal relationships which may exist between these variables. As such, statistical analysis forms a core part of this study. Use will be made of measures of central tendency (mean, median, mode), measures of dispersion (standard deviation, kurtosis, skewness, range) and various inferential statistics which will be discussed below.

2.3.1 Central Moments

This report makes use of the first four central moments (Mean, Standard Deviation, Skewness and Kurtosis). These moments form the basis of the descriptive statistics and provide valuable information on describing each samples distribution.[3]

[3]The formulas provided for each central moment are not the formulas used in the statistical analysis but instead simplified formulas for explanatory purposes

Mean

The mean ($\overline{x}$) is the first central moment of the distribution which represents the sum of a sample divided by the number of items. The calculation of the mean includes every data point and the sum of deviations from each data point to the mean is always zero. Using this statistic information such as, on average, how long an article is or how many times the word 'garda' is mentioned per article can be discovered.

$$\overline{x} = \frac{1}{N}(x_1 + \cdots + x_N) \qquad \text{(Eq. 2.3.1)}$$

Standard Deviation

The standard deviation (s) is the second central moment of the distribution. It can show how much the data deviates from the mean. A small standard deviation shows that the data tends to be close to the mean. It is calculated as being the square root of the variance. If the data is normally distributed, we can say that roughly 68% of the data lies within one standard deviation from the mean, 95% with two and 99.7% within three standard deviations.

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2} \qquad \text{(Eq. 2.3.2)}$$

Skewness

Skewness (S) is the third moment of the distribution. It measures the degree of distortion from the normal distribution. In a normal distribution, the tails on either side of the curve will be mirage images of each other and will take the iconic 'bell curve' shape. In the normal distribution the mean, median and mode will also be the same value and the level of skewness will be zero. A right (positive) skewed distribution will have a longer tail on the right and a value greater than 0.5; the greater the value, the greater the skewness. The mean will be greater than the mode and the median. A left (negative) skewed distribution will have a longer tail on the left and a value between value below -0.5; the less the value, the greater the skewness. The mean will be less than the mode and the median. A mean between -0.5 and 0.5

is considered relatively normal. If the data set of article length is positively skewed, it can be said that the many articles are below the average word count.

$$S = \frac{1}{N-1} \sum_{i=1}^{N} \frac{(x_i - \overline{x})^3}{s^3} \qquad \text{(Eq. 2.3.3)}$$

Kurtosis

Kurtosis (K) is the fourth moment of the distribution. As described by Peter Westfall "Kurtosis tells you virtually nothing about the shape of the peak - its only unambiguous interpretation is in terms of tail extremity, that is, either existing outliers (for the sample kurtosis) or propensity to produce outliers (for the kurtosis of a probability distribution)" (Westfall 2014). It, therefore, can be seen as a measure of the outliers in the distribution. A normal distribution has a Kurtosis level of 3; however, in practice "excess kurtosis" is often used which can be considered to be $Kurtosis - 3$. This makes the kurtosis of a normal distribution 0.

$$K = \frac{1}{N-1} \sum_{i=1}^{N} \frac{(x_i - \overline{x})^4}{s^4} \qquad \text{(Eq. 2.3.4)}$$

A Kurtosis of 0 is considered to be mesokurtic, any the level of kurtosis falls below this it is called platykurtic and if it is above it is considered leptokurtic. Values higher/lower than this may be due to the presence/lack of outliers or underlying data issues. Figure 2.4 visualises different levels of kurtosis, ranging from a Laplace Double exponential distribution of 3 (Red) to a normal distribution of 0 (black), to a uniform distribution of -1.2 (magenta).

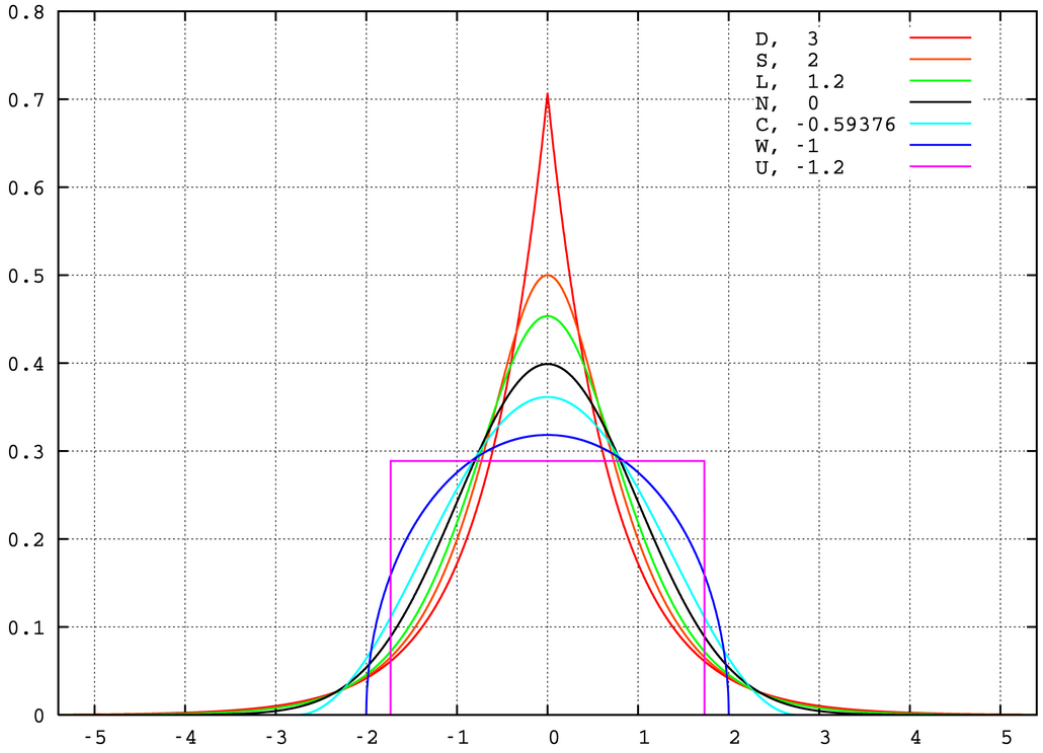

Figure 2.4: Kurtosis Visualization

2.3.2 Vector AutoRegression - VAR

Three decades ago Christopher Sims, after closely following the work of the Nobel prize-winning econometrician Jan Tinbergen, began advocating the use of VAR for forecasting economic time series and & designing and evaluating economic models (Tinbergen 1939, Sims 1980*a*,*b*, 1972). Since its inception, VAR has been used in many research projects yielding positive results (Maio & Santa-Clara 2015, Koop & Tole 2013, Pesaran 2015, Herwartz & Lütkepohl 2014). VAR's most common application is in macroeconomic analysis of financial factors such as interest rates and inflation. However, multiple papers have used VAR in tandem with sentiment analysis (Tetlock 2007, Zhao & Ahmad 2015, Ahmad et al. 2016, Kelly 2016).

The premise behind VAR, and other auto-regressive (AR) models is that past values of a variable can contain information in predicting the present (and future) value of a dependent variable. This can be represented through a univariate auto-regressive model below which contains p-lags (AR(p)).

$$y_t = c + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \ldots + \alpha_p y_{t-p} + \varepsilon_t \qquad \text{(Eq. 2.3.5)}$$

Where c is a constant, y_t is the value of the endogenous variable at lag t, α are coefficients or the model and ϵ_t is white-noise, i.e. uncorrelated over time with a constant variance and mean zero. This can be more succinctly presented as follows.

$$y_t = c + \sum_{i=1}^{p} \alpha_i y_{t-i} + \varepsilon_t \qquad \text{(Eq. 2.3.6)}$$

Often we are interested in how not only one variables lagged values effects itself, but also the lagged values of other exogenous variables. Sims (1980*b*) understanding this concern created VAR, which transformed the univariate AR process into multivariate, expressing "each variable as a linear function of its own past values, the past values of all other variables being considered, and a serially uncorrelated error term" (white noise) (Stock & Watson 2001).

Such a model lends itself to be expressed in vector and matrix notation with T observations y_0 through to y_T, p lags and k variables.

$$Y = BZ + U \qquad \text{(Eq. 2.3.7)}$$

where

$$\mathrm{Y} = \begin{bmatrix} y_p & \mathrm{y}_{p+1} & \cdots & y_T \end{bmatrix} = \begin{bmatrix} y_{1,p} & y_{1,p+1} & \cdots & y_{1,T} \\ y_{2,p} & y_{2,p+1} & \cdots & y_{2,T} \\ \vdots & \vdots & \vdots & \vdots \\ y_{k,p} & y_{k,p+1} & \cdots & y_{k,T} \end{bmatrix}$$

$$\mathrm{B} = \begin{bmatrix} c & A_1 & A_2 & \cdots & A_p \end{bmatrix} = \begin{bmatrix} c_1 & a^1_{1,1} & a^1_{1,2} & \cdots & a^1_{1,k} & \cdots & a^p_{1,1} & a^p_{1,2} & \cdots & a^p_{1,k} \\ c_2 & a^1_{2,1} & a^1_{2,2} & \cdots & a^1_{2,k} & \cdots & a^p_{2,1} & a^p_{2,2} & \cdots & a^p_{2,k} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \cdots & \vdots & \vdots & \ddots & \vdots \\ c_k & a^1_{k,1} & a^1_{k,2} & \cdots & a^1_{k,k} & \cdots & a^p_{k,1} & a^p_{k,2} & \cdots & a^p_{k,k} \end{bmatrix}$$

$$\mathrm{Z}= \begin{bmatrix} 1 & 1 & \cdots & 1 \\ y_{p-1} & y_p & \cdots & y_{T-1} \\ y_{p-2} & y_{p-1} & \cdots & y_{T-2} \\ \vdots & \vdots & \ddots & \vdots \\ y_0 & y_1 & \cdots & y_{T-p} \end{bmatrix} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ y_{1,p-1} & y_{1,p} & \cdots & y_{1,T-1} \\ y_{2,p-1} & y_{2,p} & \cdots & y_{2,T-1} \\ \vdots & \vdots & \ddots & \vdots \\ y_{k,p-1} & y_{k,p} & \cdots & y_{k,T-1} \\ y_{1,p-2} & y_{1,p-1} & \cdots & y_{1,T-2} \\ y_{2,p-2} & y_{2,p-1} & \cdots & y_{2,T-2} \\ \vdots & \vdots & \ddots & \vdots \\ y_{k,p-2} & y_{k,p-1} & \cdots & y_{k,T-2} \\ \vdots & \vdots & \ddots & \vdots \\ y_{1,0} & y_{1,1} & \cdots & y_{1,T-p} \\ y_{2,0} & y_{2,1} & \cdots & y_{2,T-p} \\ \vdots & \vdots & \ddots & \vdots \\ y_{k,0} & y_{k,1} & \cdots & y_{k,T-p} \end{bmatrix}$$

$$\mathrm{U}= \begin{bmatrix} e_p & e_{p+1} & \cdots & e_T \end{bmatrix} = \begin{bmatrix} e_{1,p} & e_{1,p+1} & \cdots & e_{1,T} \\ e_{2,p} & e_{2,p+1} & \cdots & e_{2,T} \\ \vdots & \vdots & \ddots & \vdots \\ e_{k,p} & e_{k,p+1} & \cdots & e_{k,T} \end{bmatrix}.$$

In this way the coefficient matrix B can be solved through ordinary least squares estimation of $Y \approx BZ$

For the remainder of this report VAR will be expressed in its more succinct reduced-form. A reduced-form VAR(p) can be expressed as follows

$$y_t = c + A_1 y_{t-1} + A_2 y_{t-2} + \cdots + A_p y_{t-p} + \epsilon_t \qquad \text{(Eq. 2.3.8)}$$

where p represents the current lag, c is a k-vector of intercepts, A_i is a time-invariant (k * k) matrix, e_t is a k-vector of white-noise.

Lag Selection

The length of the lag length (p) is also a critical decision that must be chosen carefully. Proper estimation of lag length is vital as discovered by Braun & Mittnik (1993), where it was discovered that lag estimates that differ from the true lag length result in inconsistent and therefore unreliable results. Further

research by Lütkepohl (2005), Hafer & Sheehan (1989) identified that overfitting results in a loss of degrees of freedom and causes an increase in the mean squared forecast error while under-fitting often resulted in auto-correlation errors. Researchers must ensure to balance having enough parameters to adequately model the relationships amongst the variables in the population (sensitivity) with ensuring the model is not overfitted or suggesting nonexistent relationships (specificity) (Dziak et al. 2019).

The first approach to lag selection is using the Akaike Information Criterion (AIC) first introduced by Akaike (1974). It estimates the relative Kullback-Leibler distance of the likelihood function of the candidate model, from the true unknown likelihood function that generated the data (Dziak et al. 2019). It can be calculated by the formula below, where k is the number of estimated parameters and $\hat{L}$ is the maximum value of the likelihood function.

$$AIC = 2k - 2\ln(\hat{L}) \qquad \text{(Eq. 2.3.9)}$$

The AIC with the lowest value is the lag which is optimal according to Akaike (1974). The $2k$ represents a penalty which is an increasing function of the number of parameters. As increasing the number of parameters will inherently improve the goodness of it, this penalty acts as a deterrent for overfitting. It is important to note that AIC only shows the relative quality of the models and not absolute quality and therefore all models could still be poor. To mitigate this risk, the absolute quality should be validated, possibly through residual testing or testing the model's predictions.

The Bayesian Information Criterion (BIC), or Schwarz criterion, is a similar criterion to AIC introduced by Schwarz et al. (1978). This criterion makes use of Bayes' theorem, determining the posterior probability of each model in order to find the optimal model. The likelihood functions between both criteria are the same; the difference is in the penalty, where BIC applies a more substantial penalty. BIC can be calculated using the equation below where n is the sample size, k is the numbers of estimated parameters and $\hat{L}$ is the maximum value of the likelihood function.

$$BIC = \ln(n)k - 2\ln(\hat{L}) \qquad \text{(Eq. 2.3.10)}$$

The final criterion considered is the Hannan-Quinn criterion (HQC) which was introduced by Hannan & Quinn (1979). As noted by Claeskens & Hjort

(2008), HQC, along with BIC, is not asymptotically efficient, unlike AIC which is. This has been further shown by Yang (2005) who states under the assumption the "true model" is not in the candidate set, AIC is asymptotically optimal for choosing the model with the least mean squared error.

$$HQC = 2k\ln(\ln(n)) - 2\ln(\hat{L}) \qquad \text{(Eq. 2.3.11)}$$

There has been much research into which criterion is superior, with a common conclusion being AIC is most accurate in selecting the true lag with small sample sizes (under 60) while HQC was seen to be superior for larger sample sizes (over 60) (Liew 2004, Ozcicek & Douglas Mcmillin 1999). BIC and HQC are preferable in some situations as they are 'consistent'. This implies that assuming the true model exists in the candidate set, a consistent model is one which is select the true model with probability approaching 100% as n $\rightarrow \infty$. The true model is the model which minimises the Kullback-Leibler distance (Dziak et al. 2019, Claeskens & Hjort 2008). Each model has its own advantage in certain situations and therefore lag selection will be carried out on a case-by-case basis. (See Asghar & Abid (2007) and Lütkepohl & Poskitt (1991) Chapter 4 for additional information on lag selection)

2.3.3 Additional Statistics

Variance

Variance represents how spread out the data is from the mean. It is calculated by averaging the squared differences from the mean and is the standard deviation squared (equation Eq. 2.3.2). The higher the variance the greater the spread of the data with a variance of 0 meaning all the data are the same value.

Z-Score

The Z-score, or standard score as it is sometimes called, is used to identify how many standard deviations above or below the mean a data point lies. It is a useful statistic in identifying outliers as it enables a comparison of the results to a normal distribution. As results are being compared to a normal distribution, it is essential to ensure the sample size is sufficiently large (>

30). If the population variance is unknown or the sample size is not sufficiently large, then the Student's t-test may be a preferable statistic. The formula is given below where x represents the raw data point being examined, $\bar{x}$ represents the sample mean and S represents the sample standard deviation.

$$z = \frac{x - \bar{x}}{S} \qquad \text{(Eq. 2.3.12)}$$

As the diagram below shows, a z-score above 0 indicates the value is above the sample mean, while a z-score below 0 indicates the value is below the sample mean.

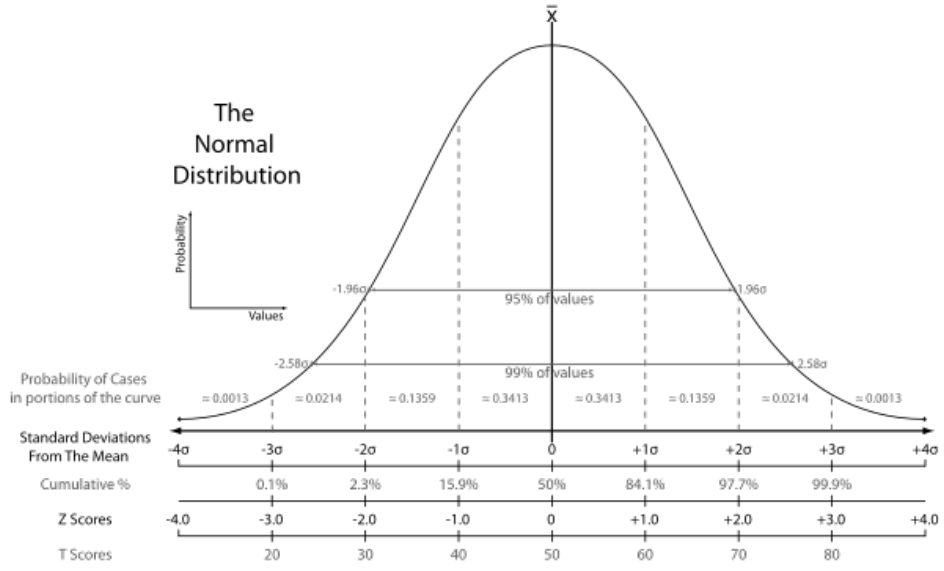

Figure 2.5: Z-score distribution

Correlation

Correlation is a statistical measurement of the linear association between two sets of values. This project makes use of the Pearson Product-Moment Cor-

relation Coefficient which ranges from +1 (Strongly Positively Correlated) to -1 (Strongly Negatively Correlated). A zero value represents no relationship between the two sets of variables. The correlation coefficient can be calculated using the formula below, where n is the sample size, x_i is the i'th indexed x value and $\bar{x}$ is the sample mean for x values.

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \qquad \text{(Eq. 2.3.13)}$$

R^2 and Adjusted R^2

R^2 is a measure of how close each data point fits to the regression line (line of best fit) and is calculated through ordinary least-squares (OLS) regression. It tells us the percentage of variation in the dependent variable which is accounted for by the regression of the independent variables. It is generally in the range of 0-1, however, values below 0 and above 1 are possible. R^2 can be calculated using the equation below (Equation Eq. 2.3.14), or if the y-intercept is estimated, it can be calculated by squaring the sample correlation coefficient between the observed values (y) and predicted data values (f) of the dependent variable (r).

$$R^2 \equiv 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}} \qquad \text{(Eq. 2.3.14)}$$

where

$$SS_{\text{tot}} = \sum_i (y_i - \bar{y})^2$$
$$SS_{\text{res}} = \sum_i (y_i - f_i)^2$$

Adjusted R^2, or $\bar{R}^2$ as it is sometimes referred to, is very similar to standard R^2 except it takes into consideration the number of independent variables used. It aims to deter a user from using 'kitchen sink regression', which is when additional, potentially irrelevant, independent variables are added to inflate the R^2 value (Barreto & Howland 2005). The Standard R^2 assumes

that the variation in the dependent variable is a result of all the independent variables, while adjusted R^2 only takes into account the independent variables whose results improve the model more than would be expected by chance alone. Adjusted R^2 will always be less than or equal to standard R^2. It can be calculated as follows where n is the sample size and p is the total number of explanatory variables.

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1} \qquad \text{(Eq. 2.3.15)}$$

P-value

The P-value, or probability value, is the probability for a given model, that the statistical summary would be greater than the actual observed results when the null hypotheses H_0 is true (Wasserstein et al. 2016).A level of significance (α) is chosen, commonly 1%, 5% or 10% representing a p-value of 0.01, 0.05 and 0.1 respectively. If the resultant p-value is below the critical level, there is strong evidence to reject the null hypotheses H_0. The VAR analysis uses significance levels of 1%, 5% and 10% while the rest of the report uses a significance level of 5% unless otherwise stated.

F-statistic

An F-test can be used when determining if a group of variables are jointly significant and therefore did not happen due to chance. The F-statistic should be used in tandem with a p-value, with H_0 stating a model with no independent variables fits the data at least as well as the model being tested. Using the F-statistic in regression can help determine if the added coefficients help improve the fit of the model to the data.

Durbin Watson

Developed by James Durbin and Geoffrey Watson, the Durbin-Watson statistic is used to test for the presence of autocorrelation in the residuals of a regression analysis at lag 1 (Durbin & Watson 1951). The test statistic is given below where e_t is the residual and t is the number of observations.

$$d = \frac{\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2} \qquad \text{(Eq. 2.3.16)}$$

One the test statistic has been determined, it can be compared against the Durbin-Watson tables to determine the presence of positive correlation. [4]. At α level significance, the test statistic should then be compared against the lower ($d_{L,\alpha}$) and upper critical values ($d_{U,\alpha}$). In this report, an α level of 5% is used. If d $>$ than $d_{U,\alpha}$ then there is no positive auto-correlation present, is d $< d_{L,\alpha}$ then there is statistical evidence of positive correlation, and if $d_{L,\alpha} <$ d $< d_{U,\alpha}$ the test is inconclusive. A rule of thumb generally applied is values of 2 indicate no auto-correlation while values less than 1 or greater than 3 represent the presence of positive or negative correlation respectively (Field & Miles 2010). This is an important test when performing regression analysis as auto-correlated residuals can lead to an underestimated standard error giving the false appearance of statistical significance.

[4] See www3.nd.edu/∼wevans1/econ30331/Durbin_Watson_tables.pdf for Durbin-Watson tables

Chapter 3

Method

This section will provide an overview of the design choices made in order to achieve the research objectives as previously outlined (Chapter 1). The steps involved in achieving these objectives can be grouped into three major categories: Data acquisition, content analysis and statistical analysis. Each stage will be discussed in detail below.

3.1 Data Acquisition

The initial section of this method relates to the process of gathering and aggregating data to create a corpus for later analysis. This section will try and achieve the research objective: *Gather a high quality and substantially large set of documents (corpus) relevant to An Garda Síochána.*

The corpus should contain the following characteristics:

Relevant Corpus

Thousands of articles are published annually by news publications worldwide. However, selecting all articles would result in irrelevant articles being gathered which would unnecessarily interfere with the statistical analysis. To combat this issue only articles with proven relevance to An Garda Síochána were selected for analysis. This criteria selected for relevant articles is the presence of the keyword 'Garda' or 'Police' anywhere in the text of an Irish news publication. A single mention of either keyword is sufficient for selection.

High Quality

There are many sources from which news could be collected. These include, but are not limited to, newspapers, Industry Trade Press, Magazines, Journals, Newswires & Press Releases, social media, and web-based publications. It is important to ensure that any source selected is of sufficient quality. Bloggers and online site have been excluded from the corpus as they did not meet the quality threshold requirements. This decision was made as these sources are generally secondary sources as identified by (Cross & Butts 2005), who have been shown to reference primarily each other and to rely on other news sources. In order to ensure the highest quality data, it was decided to focus on newspapers, magazines and journals who have made licensing agreements with LexisNexis News & Business.

Particular focus was put on ensuring that duplicate articles were not included in the corpus. It is common practice when news publications are adding to a previously made article to prepend the document with 'Update'. Care was taken to only count the most recent version of each article and discount the previous editions. Furthermore, it is also common for multiple publications to post identical articles to their viewers. This is a result of syndication and was seen to occur in the LexisNexis database, especially amongst smaller regional newspapers. This is possibly caused by a lack of competition amongst local newspapers, which Lacy (1984) found in the United States led to a decline in attention to local detail. To combat this a similarity analysis was performed on all articles and articles which have a moderate or above similarity with other articles are grouped together, with only one article from each group being selected.

Substantially Large

It is vital that the corpus gathered is substantially large as a data set too small may make the analysis susceptible to bias. There were three specific questions that had to be answered to ensure the data set was of sufficient size.

1. How many different publications should be selected?
2. When should the start date and end date for collection be?
3. How many articles should be gathered from each source?

According to the Press Council of Ireland, on a national level there are nine daily publications and seven weekly publications, 48 recognised on a local level and 13 recognised student publications (Press Council 2019). While student level publications are officially recognised, they have been omitted from the corpus as they are at present being archived in such a way that is difficult to add to the corpus or are not being archived at all. As such, a number of newspapers were selected from both the national and regional sections that archive all of their newspapers with the tool LexisNexis (Table 3.1). It is beneficial to select both national and regional publications as a concrete differential in reporting has been discovered between both sources (O'Mahony et al. 2000). It has been identified that national newspapers tend to only focus on the most serious of crime, while "local newspapers provide something akin to a court reporting service which runs the gamut from the most serious to the most mundane offences" (Black 2015). Furthermore, it has been established that local news related to crime uses less sensationalist language in its description (Healy & O'Donnell 2010). While the newspaper industry is certainly contracting, According to the Audit Bureau of Circulations (ABC) it still has a powerful readership with 2,939,714 copies in circulation in Ireland in H1 2018 (ABC 2018). This indicates there will be a sufficient quantity of articles to perform content analysis effectively.

Careful thought had the be put into the second question as sentiment differs throughout the year as different events occur. News Articles at the beginning of the week may hold a statistically different sentiment to news articles at the end of the week. Similarly, news articles at the beginning of the year may hold different sentiment to news articles at the end. Furthermore, particular events can have a large impact on the sentiment for some time, for example, the Garda whistle-blower scandal. Selecting articles within this time interval would result in results that would not accurately represent the population. To try and mitigate these effects the time range selected was from 1st January 2017 00:00 to 31st December 2018 23:59. This two-year range is deemed sufficient to mitigate the possibility of any one particular event skewing the data.

With regards to the final question, it is not necessary to take a random sample of articles from the sources as it will be feasible to collect all the articles available from the LexisNexis database. There are few publications, such as Irish News and Irish Independent, which account for the majority of articles present on LexisNexis. While over-reliance on these articles may increase potential exposure to bias, it is necessary to take all the articles available as the corpus would not be sufficiently large otherwise

Table 3.1: News Publications Chosen

Publication Name
Bray People
Corkman
Drogheda Independent
Enniscorthy Guardian
Fingal Independent
Gorey Guardian
Irish Daily Mail
Irish Examiner
Irish Independent
Irish Medical Times
Irish News
Kerryman
Louth Leader
New Ross Standard
Sligo Champion
Sunday Business Post
Sunday Independent
The Argus
The Herald
The Irish News
The Irish Times
Wexford People
Wicklow People

3.2 Content Analysis

While there are many text analysis software available, this project makes use of the RockSteady content affect analysis system, developed in Trinity College Dublin, Ireland (Ahmad 2014). RockSteady has the ability to take in a stream of structured and unstructured data, aggregate the frequency of words and according to specific categories, and further aggregate data based on a provided criteria, such as time-scales (minute, daily, monthly, yearly) or document source. As Highlighted by Ahmad et al. (2016), Rocksteady is also an automatic morphological and syntactic analysis tool, processing the ability to determine that rose, rise, risen, and rising are the morphological variants of the same root rise and that share, in a financial context, is a noun rather than

a verb. Furthermore, using RockSteady if can search for particular keywords through its GUI, which uses term frequency-inverse document frequency (TF-IDF) to identify related documents.

RockSteady uses a lexicon-based approach when performing text analysis, specifically the dictionary/bag-of-words style. In using this approach, the text is separated into token with each token being checked against a provided general and/or specialist dictionary. If a word is present in a specialist dictionary, it will update the respective category frequency and move onto the next token; If a word is not found it will continue its search in the general dictionary where if found, will update the corresponding category. The general dictionary acts as the base affect dictionary, while the specialised dictionaries are generally made to suit a specific use-case, such as the crime categories and garda people names as mentioned previously. The base dictionary used was Stones Harvard General Inquirer (GI) dictionary (Chapter 2), and six additional specialist dictionaries were provided.

3.2.1 Creating the Specialist Dictionaries

To effectively extract information specific to An Garda Síochána, crime, and the Irish judicial system, several specialist dictionaries were required. In the course of this thesis, six specialist dictionaries were created, targeting mentions of members of An Garda Síochána, mentions of Garda Departments, mentions of Garda Stations, mentions of crime, mentions of the Irish courts and mentions of their respective judges. All dictionaries were created with the aid of the corpus analysis toolkit AntConc. This provided the author with the ability to identify different word patterns used by publications as seldom did all publications use the same writing style. Creating the specialist dictionaries also ensured words with negative connotations, such as murder or extortion, were not falsely counted as a negative term in the GI dictionary as in most cases the word will be used in a descriptive sense.

Garda Departments

The Garda Department dictionary contained 13 distinct departments as per the official Garda Organisation structure (An Garda Síochána 2018*a*). In creating this dictionary, care was given to how news publications mentioned the department names, as this was not always uniform. For example, the official name for one of the departments is "Garda National Drugs & Organised

Crime Bureau" however, many publications called this either the "Garda Drugs & Organised Crime Bureau", "Garda National Drugs & Organised Crime Unit" or just the "Organised Crime unit" or "Drugs Unit". Only the top level department is incremented in the dictionary. For example, if the sub-departments the Garda National Cyber Crime Bureau, Criminal Assets Bureau, Garda National Protective Services Bureau, Garda National Bureau of Criminal Investigation, Garda National Economic Crime Bureau, Garda National Immigration Bureau or Garda National Drugs & Organised Crime Bureau appears in the text, then the Special Crime operations department is incremented as opposed to their individual counters.

Table 3.2: Garda Department Dictionary

Department Name
Security & Intelligence
Special Crime Operations
Western Region
Northern Region
Dublin Metropolitan Region
Southern Region
South Eastern Region
Eastern Region
Roads Policing & Major Event Management
Community Engagement & Public Safety
Governance & Accountability
Strategy & Transformation
Legal & Compliance
Executive Support & Corporate Services

GardaStations

This dictionary includes each garda station from Irelands 22 districts (Table 3.3). This dictionary contained all of Ireland's 564 garda stations (An Garda Síochána 2018*a*). To ensure mentions of regions were not mistakenly counted as a mention of a garda station in the area, the terms in the dictionary were listed as "'AreaName' + Garda Stations" e.g. Terenure Garda Station.

Table 3.3: Garda Station Dictionary

Region
Cavan/Monaghan
Clare
Cork City
Cork North
Cork West
Donegal
Dublin Metropolitcan
Galway
Kerry
Kildare
Kilkenny/Carlow
Limerick
Louth
Mayo
Meath
Roscommon/Longford
Sligo/Leitrim
Tipperary
Waterford
Westmeath
Wexford
Wicklow

Garda People

Table 3.4 shows the titles which were specifically targeted in this report. A core focus of this dictionary was including as many mentions as possible without double counting. It was discovered that for individuals in the Garda

force, the rank usually appeared before the name. However, it was not appropriate to use position and full names separately, as it is common for news publication to use both together, "Commissioner Drew Harris" for example. To circumvent this issue, two approaches were adopted. The first one used the title solely, e.g. any mentions of Seargent incremented the Seargent count. The second approach used the title of the individual member followed by the individual's surname. This was found to have a high level of accuracy as seldom did both approaches appear together, e.g. "Commissioner Mr Harris".

For the executive level titles, it was discovered that the title was not always present with the name [1]. For this reason, the focus was on using the full name and the title + surname, "Charlie Flanagan" and "Mr Flanagan" for example. This report focuses on the current executive committee as per the Garda Website (An Garda Síochána 2018*a*). While this report concentrates solely on articles from the years 2017/2018, there were multiple mentions of past justice ministers in such a context that was deemed worth noting. The ministers in question are Alan Shatter (2011 - 2014), Frances Fitzgerald (2014 - 2017), and Charlie Flanagan (2017 - Present).

An issue that arose when creating this dictionary is that multiple titles had the term "commissioner" in it; For example, there is the commissioner, assistant commissioner and deputy commissioner. By using just the word commissioner, RockSteady would believe that a commissioner is being mentioned any time the deputy commissioner or assistant commissioner was mentioned. To circumvent this issue, the count of commissioner mentions was calculated as total mentions of 'commissioner' − total mentions of 'deputy commissioner' − total mentions of 'assistant commissioner'.

[1] The Ministers for Justice are also included in this group

Table 3.4: Garda People Dictionary

Position
Minister for Justice
Commissioner
Assistant Commissioner
Deputy Commissioner
Chief Administrative Officer (CAO)
Chief Medical Officer
Executive Director
Director
Head of Legal Affairs
Chief Superintendent
Superintendent
Inspector
Sergeant
Garda

Courts

Table 3.5 shows the courts which were analysed under the court's dictionary. The District Court is the lowest court of which there are 24 in the Republic of Ireland. This court generally deals with summary offences, which are offences that come from legislation only and do not require a jury. This court deals with a range of criminal offences, such as speeding, drunk driving, assault and the initial hearings of more serious offences (Courts.ie 2019*a*). Due to the lower nature of this court is it most common to see Category - 13 Public Order and other Social Code Offences being tried here. Similarly, the Children Court generally takes place in the District Court building and deals with all offences concerning children under 18 except manslaughter (CitizensInformation.ie 2019). The next court up is the Circuit court of which there are 8. This court deals with all offences except for the most serious, generally being category 1 - homicide and serious offences and category 2 - sexual assault, such as rape or aggravated sexual assault. The high court deals with bail appeals and also tries the most serious offences such as those mentioned previously. When the high court exercises its criminal jurisdiction it is known as the central criminal court, therefore in the context of crime a mention of either would be referring to the same court. The special criminal court generally deals with scheduled offences such as offences against the

state or firearm/explosive offences. Therefore it is category 11 - Weapons and Explosives Offences and Category 15 - Offences against Government, Justice Procedures and Organisation of Crime which are generally tried here.

In addition to these courts there are also multiple courts which hear appeals in Ireland. The primary appeals court is the court of appeals. This court hears appeals in criminal cases from the Special Criminal Court, Central Criminal Court, or Circuit Court [2]. The highest court in Ireland is the Supreme Court, or the Court of Final Appeal as it is sometimes known. This Court hears appeals from the Court of Appeal and the High Court.

Table 3.5: Court Dictionary

Court
Supreme Court
Court of Appeal
High Court
Central Criminal Court
Special Criminal Court
Circuit Court
District Court
Children Court

[2]Since 2014, the Court of Criminal Appeal and the Courts-Martial Appeal Court were abolished and the court of appeal hears their respective cases. Therefore any mention of either court has been counted as a mention under the court of appeal.

Judges

Judges of five distinct courts were selected for this section (Table 3.6). In total 161 Judges were selected for this dictionary, whose names were retrieved from Courts Services Ireland (Courts.ie 2019*b*). The District Court has the largest number of courts and therefore the most substantial amount of judges. There are 19 moveable judges of the District Court, 25 Provincial Judges of the District Court and 16 Dublin Metropolitan Judges, including Rosemary Horgan, President of the District Court.[3] The Circuit Court has 2 Specialist Judges along with 38 Judges of the Circuit Court, including the President Mr Justice Raymond Groarke. The High Court employs 40 Judges including the President Mr Justice Peter Kelly, and the Court of Appeal employs ten judges including the President Mr Justice George Birmingham. Finally, there are 8 Judges employed in the Supreme Court.[4]

Table 3.6: Judge Dictionary

Court
Supreme Court Judge
Court of Appeal Judge
High Court Judge
Circuit Court Judge
District Court Judge

[3]At the time of writing there is 1 Dublin Metropolitan Judge vacancy and 1 Provincial Judges of the District Court vacancy

[4]At the time of writing there is two vacancies in the Supreme Court

Crime

A crime is recorded when it is reported to a member of An Garda Síochána, and upon assessing the evidence believes that a criminal offence by law has occurred. The process of reporting crime has changed drastically in Ireland. Throughout the 19th Century, official crime statistics were centred around the counting of court proceedings. This had a major drawback as only a fraction of crimes make it to trial. In the 20th Century, the criminology field made major advancements, one, in particular, was the research of Thorsten Sellin who published "The Basis of a Crime Index" in Fall 1931. In this, he concluded that "the value of a crime rate for index purposes decreases as the distance from the crime itself in terms of procedure increases" (Sellin 1931). Unsurprisingly, Ireland switched to counting the number of reported crimes in 1864, a system which remains today. The Irish Crime Categorisation System (ICCS) is the present name for the system and was developed by the CSO (CSO 2019). At the time of writing Version 2 is the most recent version and was introduced on 05/01/2017.

Table 3.7: Crime Dictionary

Category	Title
1	Homicide
2	Sexual Offences
3	Attempts/Threats to Murder, Assaults, Harassments and Related offences
4	Dangerous or Negligent Acts
5	Kidnapping and Related Offences
6	Robbery, Extortion and Hijacking Offences
7	Burglary and Related offences
8	Theft and Related Offences
9	Fraud, Deception and Related Offences
10	Controlled Drug Offences
11	Weapons and Explosives Offences
12	Damage to Property and to the Environment
13	Public Order and other Social Code Offences
14	Road and Traffic Offences (NEC)
15	Offences against Government, Justice Procedures and Organisation of Crime
16	Offences Not Elsewhere Classified

3.3 Statistical Analysis

3.3.1 Vector AutoRegression (VAR) Analysis

VAR is used to identify any causal relationships which may exist between the variables. The general VAR equation for P-lags is seen as follows:

$$y_t = c + A_1 y_{t-1} + A_2 y_{t-2} + \cdots + A_p y_{t-p} + e_t \qquad \text{(Eq. 3.3.1)}$$

The GNU Regression, Econometrics and Time-series Library (GRETL) was used for the purpose of VAR analysis. GRETL is an open-source statistical package written in C providing both a GUI in GTK+ and has a command-line interface. It has powerful graphing capabilities, calling gnuplot upon requests, an also has the ability to perform additional statistical analysis such as cointegration tests, ANOVA, and Least Squares Regression. GRETL was chosen as the statistical tool as its performance and accuracy has been verified by many journals, such as the Journal of Statistical Software and The Journal of Applied Econometrics (Baiocchi & Distaso 2003, Rosenblad 2008, Yalta & Yalta 2007, Mixon Jr & Smith 2006).

A VAR analysis can be performed in GRETL by providing at least one endogenous variable and any additional number of endogenous/exogenous variables. A variable is considered to be endogenous if its value is determined by the states of other independent variable is the causal model. A fully endogenous variable is one whose value is wholly determined by other variables. On the other hand, an exogenous variable is one whose value is determined independently of the states of other variables in the model and, therefore, can be considered to be fully causally independent from the system.

It is important to ensure the variables are stationary before performing VAR, which means its statistical properties, e.g. mean, variance, auto-correlation are standard over time. Each variable can be tested for stationarity by performing a unit-root test. This report makes us of the Augmented Dickey-Fuller (ADF) Test to test for unit-roots (Fuller 2009). The null hypothesis states that the unit-root is present while the alternate hypotheses, depending on the version of the test used, states the time-series sample has stationarity or trend-stationarity. An ADF test is applied to the following model:

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \delta_1 \Delta y_{t-1} + \cdots + \delta_{p-1} \Delta y_{t-p+1} + \varepsilon_t, \qquad \text{(Eq. 3.3.2)}$$

Where α is a constant, β is the coefficient on the time-trend, and p is the lag order. It is required to determine the optimal lag length prior to running this model. There are multiple ways to calculate the optimal lag length (Chapter 2), for the ADF test the Akaike information criterion (AIC) criterion is used. Once the Test Statistic is computed it is compared against the corresponding critical value (Table 3.8). If the Test Statistic is smaller (negative) than the critical value, the null hypothesis is rejected.

Table 3.8: Dickey-Fuller Critical Values for t-distribution (Fuller 2009)

	With Trend		Without Trend	
Sample size	1%	5%	1%	5%
T = 25	−3.75	−3.00	−4.38	−3.60
T = 50	−3.58	−2.93	−4.15	−3.50
T = 100	−3.51	−2.89	−4.04	−3.45
T = 250	−3.46	−2.88	−3.99	−3.43
T = 500	−3.44	−2.87	−3.98	−3.42
T = ∞	−3.43	−2.86	−3.96	−3.41

If the above test failed, this indicates that there is a unit-root present in the variable and therefore is non-stationary. It is important to next test for cointegration between the variables using the Engle-Granger cointegration test (Engle & Granger 1987). In this test, a unit-root test is performed on the residuals of the cointegration test, if the residual has no unit-root the series is considered cointegrated and the Vector Error Correction Model (VECM) should be used rather than VAR. The null hypothesis states the are not cointegrated and the critical values as presented by MacKinnon (1991) must be used.

If the original ADF test and cointegration test both fail, then the variables must be made stationary. It is important to convert any non-stationary variable to stationary as they can provide spurious regression results in multivariate frameworks. The process of changing the variable to stationary depends on whether it is a trend, cycle, random walk or is some combination of the three. For a difference-stationary process, it can be made stationary by differencing (Equation Eq. 3.3.3). Cointegration is preferable to differencing as the latter throws away some long-term properties of the data. A non-stationary series that is required to be differenced d times until it is stationary is considered to be 'integrated of order d', which is denoted as y_t $I(d)$. Otherwise, if the process is trend-differenced, then it can be removed by fitting an applicable model, e.g. linear or polynomial, to the data.

This regression curve can be removed from the original data which detrends it by only leaving the residuals.

$$(1 - L)Y_i = Y_i - Y_{i-1} \qquad \text{(Eq. 3.3.3)}$$

Once all variables are fully stationary, a VAR analysis can be performed. Before any real relationship can be identified a granger-causality test is performed on all variables to determine the direction of causality. Granger provides a definition of a causal relationship, stating "a (time-series) variable A causes B, if the probability of B conditional on its own past history and the past history of A does not equal the probability of B conditional on its own past history alone" (Granger 1980, 1969). Granger Causality is a popular causality test in the econometrics field, with Hoover stating it is "perhaps the most influential explicit approach to causality in economics" (Hoover 2008).

Additional variables are added into the model incrementally to test if it improves goodness of it and uncovers any underlying causal relationships. The Durbin-Watson statistic is checked each iteration to ensure to sign of auto-correlation (serial correlation) is present. Auto-correlation errors are well studied, with Granger et al. (2001) identifying three major consequences of their use.

- Estimates of the regression coefficients are inefficient.
- Forecasts based on the regression equations are sub-optimal.
- The usual significance tests on the coefficients are invalid.

Models Selected

In the creation of a VAR panel, multiple models are specified whereby initially a VAR model is performed on the variable of interest itself, afterwards adding additional variables in a step-wise fashion. After each test was performed, each model was analysed for statistical significance and fit for the data using the statistical methods outlined previously (Chapter 2).

Each model is tested for Granger-causality as discussed previously. A lack of Granger-causality in the direction of the dependent variables results in that variable being omitted from future models as the variable is shown to

have no causal effect whatsoever. In the presence of bi-directional causality or granger-causality in the direction of the dependent variable, the added variable will remain in the model for future iterations.

The models presented below highlight the most effective variables found throughout the course of this thesis. For all models, all different combinations of variables have been tested, including all sentiment variations and variables relating to the Garda specialist dictionaries with the most performant models highlighted below.

Consumer Sentiment Index

This model aims to identify if there are any causal relationships between the Irish news media and the Consumer Sentiment Index which can be viewed as an overall judge of a countries economy as judged by the opinion of the consumers. In a paper by O'Connell (1999), evidence of a "top-down" relationship between the media and sentiment of the public was identified; however, the authors were unable to verify it conclusively.

As the consumer sentiment index is only available in monthly form, all other variables have to conform to this time period. The VAR begins by regressing the Consumer Sentiment Index against itself to see if past values have any statistical relationship to present values.

$$CSI_t = const + \alpha_1 CSI_{t-1} + \alpha_2 CSI_{t-2} + \alpha_3 CSI_{t-3} + \cdots + \alpha_p CSI_{t-p} + \epsilon_t \quad \text{(M1a)}$$

where c represents the constant, p represents the numbers of lags, t represents the current time period, α represents the coefficient and ϵ represents white-noise. This formula can be displayed more succinctly as follows and will continue to be represented this way for the remainder of the study.

$$CSI_t = const + \alpha L_p CSI_t + \epsilon_t \quad \text{(M1a)}$$

Variables are then added incrementally to see if the fit of the model to the data improves, beginning with Crime. It is expected that crime and CSI will be negatively correlated, as discussion of crime goes up the public begin to believe that there is an increased breakdown of law & order in society. This can be represented by the following formula.

$$CSI_t = const + \alpha L_p CSI_t + \beta L_p Crime_t + \epsilon_t \quad \text{(M1b)}$$

Following on, negative sentiment in the news media was added as an exogenous variable. Identification of a causal relationship after the addition of this variable would indicate that the media has a quantitative effect on how the public perceive the country as a whole is performing.

$$CSI_t = const + \alpha L_p CSI_t + \beta L_p Crime_t + \lambda L_p Negative_t + \epsilon_t \quad \text{(M1c)}$$

Finally, citations of Garda people is added to the model. It is believed that in times of economic prosperity i.e. when the consumer sentiment is high, that the level of serious crime is lower. Due to the fact the Gardaí are generally mentioned in the context of crime, this could imply an inverse relationship between citations of Garda people and the CSI. This provides us with the final model 1(d) below.

$$CSI_t = const + \alpha L_p CSI_t + \beta L_p Crime_t + \lambda L_p Negative_t + \rho L_p GP_t + \epsilon_t \quad \text{(M1d)}$$

Garda People

The models choose citations of Garda People as its endogenous variable. This model focused on a daily period as over a long period of the fluctuations average out. This model aimed to identify which variables increased the discussion of Garda People in the news media. A VAR analysis is performed firstly on the variable itself which can be represented as follows:

$$GP_t = const + \alpha L_p GP_t + \epsilon_t \quad \text{(M2a)}$$

Following this test, additional variables are added incrementally to see if it will improve the fit of the model to the data. Model 1b adds Negative terms to the model to test the relationship between negative sentiment and mentions of An Garda Síochána. Hostile terms, a subset of Negative terms will also be tested to see if it provides a better fit to the data.

$$GP_t = const + \alpha L_p GP_t + \beta L_p Neg_t + \epsilon_t \quad \text{(M2b(i))}$$
$$GP_t = const + \alpha L_p GP_t + \beta L_p Hos_t + \epsilon_t \quad \text{(M2b(ii))}$$

The next variables added the to model was mentions of crime. This is expected to have a strong positive relationship with the dependent variables as crime and An Garda Síochána are usually discussed in the same articles.

$$GP_t = const + \alpha L_p GP_t + \beta L_p Neg_t + \lambda L_p Crime_t + \epsilon_t \quad \text{(M2c)}$$

The following variable added in mentions of the Garda Stations. Similarly to mentions of crime, it is expected that this variable will have a strong positive relationship to the dependent variable as news publications tend to mention both in the same articles.

$$GP_t = const + \alpha L_p GP_t + \beta L_p Neg_t + \lambda L_p Crime_t + \rho L_p GS_t + \epsilon_t \quad \text{(M2d)}$$

The final variable added to the model is mentions of judges. As the arresting garda office can be present in the courtroom and the hearing it is possible that these variable will have a strong relationship. Model 1e below represents the final fully formed model

$$GP_t = const + \alpha L_p GP_t + \beta L_p Neg_t + \lambda L_p Crime_t + \rho L_p GS_t + \gamma L_p Judge_t + \epsilon_t \quad \text{(M2e)}$$

Negative Sentiment

The final model of interest investigates how which variables effect the level of negative sentiment used by the news media. This model focused on a daily period as over a long period of the the fluctuations average out. Firstly negative sentiment is tested against lagged versions of itself which can be expressed below (model 2a).

$$Neg_t = const + \alpha L_p Neg_t + \epsilon_t \quad \text{(M3a)}$$

Following this four additional variables are tested incrementally to see how they relate to negative sentiment in the news media. The variables added are mentions of Garda People, crime, Garda Stations and Courts and can be expressed in the models below (Model 3b $\rightarrow$ Model 3e).

$$Neg_t = const + \alpha L_p Neg_t + \beta L_p GP_t + \epsilon_t \quad \text{(M3b)}$$

$$Neg_t = const + \alpha L_p Neg_t + \beta L_p GP_t + \lambda L_p Crime_t + \epsilon_t \quad \text{(M3c)}$$

$$Neg_t = const + \alpha L_p Neg_t + \beta L_p GP_t + \lambda L_p Crime_t + \rho L_p GS_t + \epsilon_t \quad \text{(M3d)}$$

$$Neg_t = const + \alpha L_p Neg_t + \beta L_p GP_t + \lambda L_p Crime_t + \rho L_p GS_t + \gamma L_p Court_t + \epsilon_t \quad \text{(M3e)}$$

Chapter 4

Experiments & Evaluation

The previous chapter laid the roadmap to achieving the research objectives which were set previously (Chapter 1). Firstly, this chapter will begin with a discussion of the corpus collected, analysing each source uniquely. Following this content analysis was performed on the corpus, including sentiment analysis, identifying the emotions used by newspapers as well as the frequency of term mentions from the specialist dictionaries created. Finally, vector autoregression is performed on the models as mentioned previously (Chapter 3) to identify if any statistically significant relationships exist between the variables.

4.1 Data Collection

Followed the criteria outlined previously (Chapter 3) a news corpus was collected from LexisNexis News & Business and is displayed below (Table 4.2). In total, over the two years, 41,779 articles were gathered from 23 sources. The bulk of the corpus is taken up by The Irish Times, Irish Independent and Irish Daily mail with 8364, 7999 and 7024 respectively, accounting for just over 56% of the entire corpus. This can be seen from the positive skew shown in table 4.1. In total, 729 days[1] were accounted for resulting in an average of 57.3 articles per day. The article length from each source was also analysed, with the average length being 560 words with a standard deviation of 221. Sunday Business Post was shown to produce the longest articles on average

[1]Data was collected every day, except for February 5th 2017 as no articles were present on the LexisNexis Database. To avoid interfering with the statistics or possible presenting data quality issues from manual entry, this day was omitted from the statistics. Resulting in February 2017 having 27 days and therefore 2017 having 364 days. This is consistent with Tetlock et al. (2008) who excluded all dates on which there were no news articles

at 1243 words per article while the Irish Medical Times had the shortest on average at 298 words.

Table 4.1: Article Summary Statistics

	Mean	St Dev	Skewness	Kurtosis	Minimum	Maximum
Articles	1816	2639	1.68	1.67	22	8364

Table 4.2: Article Breakdown by Source

Source	2018	2017	Total	Terms	Mean words per article
Bray People	92	39	131	68446	522
Corkman	107	182	289	213170	738
Drogheda Independent	103	231	334	157208	471
Enniscorthy Guardian	63	13	76	43972	579
Fingal Independent	121	48	169	106183	628
Gorey Guardian	68	22	90	59257	658
Irish Daily Mail	3464	3560	7024	4035145	574
Irish Examiner	1047	1752	2799	1563786	559
Irish Independent	3623	4376	7999	4186831	523
Irish Medical Times	16	6	22	21277	967
Irish News	3158	1075	4233	1581602	374
Kerryman	105	253	358	180821	505
Louth Leader	589	635	1224	364834	298
New Ross Standard	51	22	73	45362	621
Sligo Champion	133	307	440	238526	542
Sunday Business Post	250	506	756	939776	1243
Sunday Independent	914	1130	2044	2102921	1029
The Argus	175	37	212	91085	430
The Herald	1076	41	1117	507030	454
The Irish News	85	3349	3434	1228383	358
The Irish Times	3934	4430	8364	5312271	635
Wexford People	62	379	441	286275	649
Wicklow People	115	35	150	70852	472

4.2 Content Analysis

This section provided an analysis of sentiment and mentions of the six categories as described in Chapter 3. Table 4.3 provides an overview of the correlation discovered between the variables where bolded values represent

a statistical significance of $P < 0.05$. Negative sentiment shows a strong positive correlation with Weak, Political, Power, Passive and Crime terms. Mentions of Garda Stations was found to be negatively correlated to negative terms at -48%. Mentions of Judges and Crime are strongly negatively correlated to Positive terms while Economic terms were positively correlated. As expected, crime is negatively correlated with positive words and positively correlated with hostile and negative terms, however, it was also discovered that passive and weak terms are correlated with crime. An increase it political terms is negatively correlated will nearly all categories related to An Garda Síochána. A similar trend can be seen with strong and power terms which may be an indication of the opinion of the media with regards to An Garda Síochána.

Sentiment Analysis

Table 4.4 provides a breakdown of the relative sentiment from each source. The table indicates that Active and Strong terms account for the largest proportion of terms used. The small standard deviations indicate that for nearly all sources it accounts for roughly between 6% and 7%. Positive words appear to account for a higher proportion of terms used than Negative terms. A possible reason for this was the inclusion of Irish Medical Times which has 22 articles and a relative positive percentage of 5.43% which resulted in a kurtosis of 8.2. Positive(b) in the table below excludes Irish Medical Times from the data set, however, while kurtosis dropped to a more reasonable 1.8 and skewness became inverted, the mean percentage of positivity only dropped by 0.09%. This shows that news publications tend to use more positive words than negative words when discussing matters related to An Garda Síochána.

This occurrence may be explained by the Pollyanna hypothesis which states that there is a universal human tendency to use more positive terms than negative terms. This hypothesis was tested by Devitt & Ahmad (2013) in a similar study to this. Four dictionaries were analysed, the General Inquirer dictionary, Dictionary of affect in language (Whissell 1989), WordNet affect (Strapparava et al. 2004), and SentiWordNet (Esuli & Sebastiani 2006) and found all lexicons had a significant positive polarity bias. When used against the British National Corpus (BNC), a corpus of over 100m words, the GI dictionary was found to have the most extreme polarisation with a positive to negative ratio of 1:0.64. This is a surprising discovery considering that the ratio of positive to negative terms in the dictionary itself is 1:0.84. Fur-

Table 4.3: Correlation Matrix of Sentiment and Garda Variables

	Terms	Active	Econ	Hostile	Milit	Negative	Polit	Passive	Positive	Power	Strong	Weak	Crime	GdaPpl	Judge	Dpt	GdaStat	Court
Terms	100	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Active	-15	100	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Econ	3	3	100	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Hostile	35	**43**	-20	100	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Milit	9	32	-7	32	100	-	-	-	-	-	-	-	-	-	-	-	-	-
Negative	**45**	39	-3	**91**	16	100	-	-	-	-	-	-	-	-	-	-	-	-
Polit	**55**	17	35	**44**	22	**63**	100	-	-	-	-	-	-	-	-	-	-	-
Passive	41	7	13	**49**	15	**72**	**70**	100	-	-	-	-	-	-	-	-	-	-
Positive	-22	-41	**43**	**-70**	-18	**-56**	0	-23	100	-	-	-	-	-	-	-	-	-
Power	34	37	27	**49**	16	**67**	**90**	**69**	-4	100	-	-	-	-	-	-	-	-
Strong	8	23	**45**	6	1	27	**71**	34	39	**81**	100	-	-	-	-	-	-	-
Weak	31	29	10	**54**	-15	**79**	**67**	**80**	-17	**75**	**47**	100	-	-	-	-	-	-
Crime	8	38	-18	**74**	10	**72**	0	**49**	**-75**	35	-9	**43**	100	-	-	-	-	-
GardaPeople	4	-2	-12	-9	-13	-15	**-49**	-25	-30	**-52**	**-67**	-28	18	100	-	-	-	-
Judge	-22	41	-19	20	-14	7	**-59**	-41	**-42**	**-42**	**-46**	-15	32	**57**	100	-	-	-
Department	-14	-16	-16	7	12	1	-40	2	-31	-37	**-55**	-18	36	**65**	37	100	-	-
GardaStation	**-45**	-13	-36	-30	4	**-48**	**-78**	**-43**	-17	**-76**	**-79**	**-58**	9	**60**	41	**65**	100	-
Court	-7	36	-14	29	-5	15	**-44**	0	-33	-29	-39	-5	23	**45**	**89**	26	20	100

Bolded values represent a statistical significance of $p < 0.05$

thermore, this ratio became more pronounced when compared against more domain-specific corpora, with the BNC Informative sub-corpora receiving a ratio of (1:0.41) and financial news-test corpora receiving a ratio of (1:0.56). A similar finding was found by Kennedy & Inkpen (2006) who also used the GI dictionary.

While the proportion of negative terms is less than the proportion of positive terms, compared to finding by Cook et al. (2016) it appears the proportion of negative terms is still relatively high. In this research paper six corpora were analysed, the Corpus of Contemporary American English (COCA) and British National Corpus (BNC) to represent the American and British English respectively, and four specialist corpora related to financial news with the General Inquirer used as the dictionary. Cook et al. (2016) discovered that the mean negative terms frequency across all the corpora was 2.14%, 0.84% less than what was found in this study. However, unlike Cook et al. (2016) who discovered negative sentiment tends to have higher kurtosis and skewness than would be expected in a normal distribution, this report found the skewness to be relatively normal and kurtosis negative. This report also found the negative sentiment to be normally distributed based off the Shapiro-wilk test at a level of $p < 0.001$ whereas Cook et al. (2016) discovered all but two cases rejected the null hypothesis of normality.

Table 4.4: Sentiment recorded from total corpus

	Mean	St Dev	Skewness	Kurtosis	Minimum	Maximum
Positive(a)	3.58%	0.51%	2.0	8.2	2.61%	5.43%
Positive(b)	3.49%	0.32%	-0.9	1.8	2.61%	4.00%
Negative	2.98%	0.39%	0.3	-0.7	2.43%	3.72%
Active	6.14%	0.23%	0.5	-0.4	5.79%	6.67%
Passive	2.62%	0.28%	-0.1	-0.1	2.11%	3.25%
Strong	6.62%	0.45%	0.9	0.9	6.07%	7.81%
Weak	1.76%	0.18%	-0.6	-0.2	1.44%	2.13%
Econ	1.06%	0.17%	7.7	2.3	0.85%	1.69%
Polit	2.26%	0.56%	0.7	-0.9	1.55%	3.38%
Hostile	1.61%	0.24%	1.3	1.3	1.30%	2.22%
Power	2.62%	0.46%	1.2	1.1	2.06%	3.85%
Milit	0.23%	0.07%	1.1	1.7	0.12%	0.42%

Positive(b) removes Irish Medical Times from the data which has a positivity of 5.43% but only 22 articles.

Crime Mentions

It was found that mentions of the 16 categories of crime accounted for on average 0.183% of all terms with a standard deviation of 0.083%. The top three publications with the highest mentions of crimes are The Herald, The Argus and The Louth Leader with the relative percentages of 0.3968%, 0.3403%, and 0.3021%. The Herald is unsurprisingly the publication with the largest relative mentions of crime as this is a common finding by studies since early 1980's when the Herald began prioritising crime news in an attempt to resonate with young, urban professionals (Kerrigan & Shaw 1985, Dahlgren & Sparks 1992). There is no clear distinction between whether local/regional or national newspapers discuss crime more often with the top 5 publications for crime mentions consisting of three national newspapers and two regional. The findings also appear to back up Black (2015) who stated: "while national titles focus on the most serious crimes, local newspapers provide something akin to a court reporting service which runs the gamut from the most serious to the most mundane offences". Looking a the most "mundane" categories such as category 13 (Public Order and other Social Code Offences), it is seen that regional newspapers account for 6 of the top 7 publications. In comparison, with more serious offences such as category 11 (Weapons and Explosives Offences) 7 out of the top 8 publications are national newspapers.

Table 4.5 shows the category breakdown of crime mentioned with category 1 (Homicide), 2 (Sexual Offences), and 11 (Weapons & Explosives Offences) accounting for the largest proportion with 27.36%, 14,58% and 11.54%. Table 4.5 also shows a breakdown of offences reported to An Garda Síochána throughout 2017 and 2018 with category 1 (Homicide), 5 (Kidnapping & Related Offences), and 6 (Robbery, Extortion & Hijacking Offences) accounting for the smallest proportion with 0.04%, 0.05% and 1.08% respectively. Table 4.5 also provides a ratio of the proportion of crime recorded to the proportion of crime mentioned in the corpus. A relationship can be seen that the more frequent a crime occurs in society, the less it appears in the media, for example, category 1 & 2 have the lowest proportion recorded yet have the two highest ratios with category 1 earning a score of 684 and category 2 a score of 35.83. This ratio can be considered to be a measure of 'sensationalism' as used by journalists. These findings are akin to the results discovered by O'Connell (1999) 20 years ago in the analysis of Irish newspapers.

Table 4.5: Ratio of crime mentions vs crime statistics

Category	2017	2018	Total	Proportion Recorded	Proportion Mentioned	Ratio
1	83	74	157	0.04%	27.36%	684
2	2884	3182	6066	1.42%	14.58%	10.27
3	18925	19955	38880	9.08%	9.98%	1.1
4	8374	8553	16927	3.95%	8.01%	2.03
5	129	128	257	0.06%	2.15%	35.83
6	2186	2432	4618	1.08%	4.17%	3.86
7	19182	16969	36151	8.44%	3.88%	0.46
8	69283	67127	136410	31.86%	4.46%	0.14
9	5436	6434	11870	2.77%	11.54%	4.17
10	16792	18390	35182	8.22%	0.40%	0.05
11	2377	2434	4811	1.22%	0.25%	0.21
12	23209	21533	44742	10.45%	2.86%	0.27
13	31199	31990	63189	14.76%	4.00%	0.27
14	N/A	N/A	N/A	N/A	0.67%	N/A
15	13740	15173	28913	6.75%	2.73%	0.40
16	N/A	N/A	N/A	N/A	2.95%	N/A

No data was present for Category 14 and 16 from the CSO
Source: https://www.cso.ie/en/releasesandpublications/ep/p-rc/recordedcrimeq42018/

Garda People

Garda People are seen to be mentioned on average 4116 times per source over the past two years. In relative terms, garda people citations account for 0.395% of all terms with 0.171% standard deviation. Garda People appear to be mentioned more prominently in regional newspapers with five of the top six being regional, although the publication with the highest proportion is the Irish Examiner, a national newspaper with a proportion of 0.7695%. Table 3.4 provides a breakdown of the citations of Garda People, with general Garda personnel, the Commissioner, and Sergeants taking up the biggest proportion with 59.57%, 14.93% and 11.28% respectively. A trend can be seen in that higher level personnel such as the Minister for Justice and Equality and the Garda Commissioner receive the highest proportion of citations from national publications with the top five in each all being national publications. Furthermore, it appears general level operatives such as general Gardaí, Sergeants, and Inspectors receive more citations from regional newspapers with the Gorey Guardian, The Argus, and the Sligo Champion providing the most mentions respectively.

Table 4.6: Garda People Analysis

Title	Proportion
Garda	59.57%
Commissioner	14.93%
Sergeant	11.28%
Minister	5.66%
Superintendent	2.83%
Inspector	2.67%
Assistant Commissioner	1.26%
Chief Superintendent	0.91%
Deputy Commissioner	0.42%
Executive Director	0.34%
Head of Legal	0.08%
CAO	0.03%
Director	0.02%
CMO	0.00%

Garda Stations and Departments

Departments of An Garda Síochána received 2850 mentions in the two years analysed. The Special Crime Operations department accounted for the most substantial proportion with 2202 mentions or 77.26%. This was followed by Security & Intelligence department and Dublin Metropolitan Region department with 8.25% and 2.81% respectively. This is a relatively expected result as the Special Crime Operations department not only has the most significant amount of sub-departments but also contains most of the departments that would be mentioned with sensationalist journalism. Such departments include the Cyber Crime Bureau, Criminal Assets Bureau, National Immigration Bureau, Bureau of Criminal Investigation, Economic Crime Bureau and the Drugs and Organised Crime Bureau.

Table 4.7: Garda Department Analysis

Department	Absolute Mentions	Proportion of total
Special Crime Operations	2202	77.26%
Security & Intelligence	235	8.25%
Dublin Metropolitan Region	80	2.81%
Southern Region	76	2.67%
Western Region	54	1.89%
Northern Region	52	1.82%
Roads Policing & Major Event Management	49	1.72%
Eastern Region	45	1.58%
South Eastern Region	25	0.88%
Governance & Accountability	14	0.49%
Community Engagement & Public Safety	12	0.42%
Legal & Compliance	3	0.11%
Strategy & Transformation	2	0.07%
Executive Support & Corporate Services	1	0.04%

Similarly, Garda Stations were mentioned 4359 times throughout the two years. Garda stations in the Dublin Metropolitan area received the largest amount of mentions accounting for 40.56% of all mentions, followed by Louth with 8.56% and Wicklow with 5.46%. Mentions of Garda Stations were also grouped together into their regions in figure 4.9 by summing together the respective regional divisions from table 4.8. This table shows that the Dublin Metropolitan Region accounts for the most significant proportion of mentions with 40.56% while the Western region received the smallest number of mentions with 5.48%.

Table 4.8: Garda Station Mentions Breakdown

Region	Absolute Mentions	Proportion of total
Dublin Metropolitan	1768	40.560%
Louth	373	8.557%
Wicklow	238	5.460%
Sligo/Leitrim	210	4.818%
Wexford	194	4.451%
Kerry	188	4.313%
Cork City	186	4.267%
Cork North	133	3.051%
Donegal	130	2.982%
Kildare	112	2.569%
Cork West	105	2.409%
Limerick	103	2.363%
Clare	94	2.156%
Cavan/Monaghan	90	2.065%
Meath	76	1.744%
Kilkenny/Carlow	76	1.744%
Waterford	68	1.560%
Mayo	66	1.514%
Tipperary	45	1.032%
Roscommon/Longford	42	0.964%
Galway	37	0.849%
Westmeath	25	0.574%

Table 4.9: Regional Garda Station Mentions

Region	Proportion of total
Dublin Metropolitan	40.56%
Eastern	10.35%
Northern	18.42%
Western	5.48%
South Eastern	8.79%
Southern	16.40%

Figure 4.1 displays the each region as well as the proportion of crime which occurred in that region across 2017/2018 according to the CSO crime statistics. By comparing the regional distribution of crime with the regional mentions of Garda stations in the Irish news media some striking similarities can be

seen, with all but one of the regions differing by less than 3%. This could possibly indicate the mentions of Garda Stations in the news papers could be a reliable proxy for crime levels throughout the country although this can not be said for certain from the given analysis.

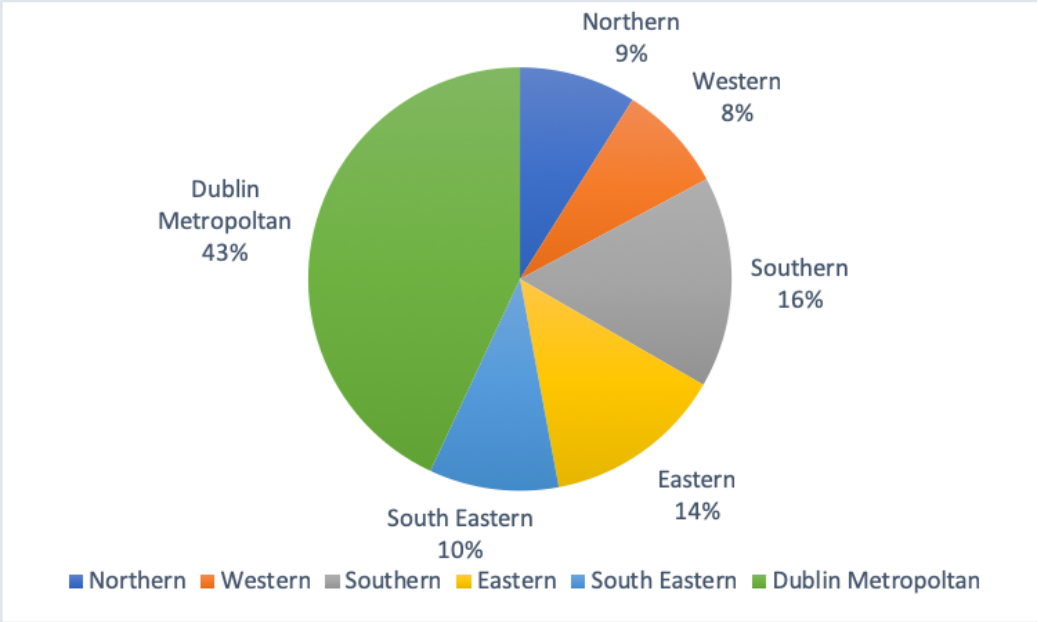

Figure 4.1: Regional Crime Distribution 2017-2018
Source: CSO Crime Statistics

Courts and Judges

The district court was the most common court resulting in 44.37% of the total court mentions. This is surprising considering the bulk of the cases the district court deals with are summary offences which make up a small proportion of the total crime mentioned in the newspapers analysed. A possible explanation for this is that the court sometimes deals with the initial hearing of more serious offences and offences such as speeding and drunk driving (Elder et al. 2004). The high court received the second highest amount of mentions accounting for 22.19% of all mentions. This large proportion is largely due to the fact that the person charged with murder can only apply to the high court for bail. The Supreme Court, was expected to have a high proportion of mentions; however it was found to have a similar proportion to the Circuit Court with roughly 8.5% each. The Central Criminal Court

and Special Criminal Court, who deal with cases such as murder, rape and terrorism had a relatively low amount of mentions considering how sensationalist the crimes are. In total both of these courts accounted for only 8% of all mentions. It was expected for regional publications to cite the lower courts most, such as the District court, while national newspapers cited the higher courts most, such as the criminal, circuit or supreme court. This was shown to not be the case and no clear distinction could be made. The Irish Medical Times was seen to cite the courts the most, especially the district court mentions it twice as much as the second most publication.

Table 4.10: Court Analysis

Court	Absolute Count	Proportion of total
Childrens Court	23	0.19%
Circuit Court	969	8.19%
Supreme Court	1050	8.87%
District Court	5249	44.37%
High Court	2744	23.19%
Special Criminal Court	366	3.09%
Central Criminal Court	595	5.03%
Court of Appeals	835	7.06%

Table 4.11 displays the results of the judges which have been mentioned in the corpus analysed. The mentions of judges are comparable to the mentions of their respective courts. District court judges also accounted for the largest proportion at 40.99%. Similarly to the courts, the Circuit Court Judges and Supreme Court Judges had a very similar proportion of mentions with just over 19%. The High court had the most disparate result with judges being mentioned four and a half times less than the court itself. Similarly to the citations of courts, there is no distinction between regional and national publications in the mentions of judges.

Table 4.11: Judge Analysis

Judge	Absolute Count	Proportion of total
Circuit Court	782	19.26%
Supreme Court	774	19.06%
District Court	1664	40.99%
High Court	632	15.57%
Court of Appeal	208	5.12%

4.3 VAR Analysis

4.3.1 Consumer Sentiment Index

This section performs a VAR analysis on Consumer Sentiment Index, adding various variable gathered from the corpus into the model to test if any meaningful relationship exists. As discussed in the previous section, VAR is performed by first regressing the endogenous variable against itself, followed by the incremental addition of further exogenous or endogenous variables. However, this was not possible in this case as the Consumer Sentiment Index is only available in monthly form, therefore, all other variables had to be converted into a monthly time-series. This resulted in a reduction of observations, so much so that a VAR with more than two variables due to lack of observations causing spurious results.

Firstly all variables were tested for stationarity through the use of the Augmented Dickey-Fuller (ADF) test. Using the critical values as outlined previously (Chapter 3) there were no unit-roots identified in each variable, and therefore all variables were confirmed as being stationary.

Subsequently, the lag length (p) was determined through the combinatory use of Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Hannan-Quinn criterion (HQC). The results of this analysis can be seen in table 4.12 were lag 7 was collectively selected as being the ideal lag.[2]

Table 4.12: Model 1 Lag Determination

lags	AIC	BIC	HQC
1	20.840579	21.134655	20.869811
2	20.967428	21.457553	21.016147
3	21.236722	21.922897	21.304929
4	21.329827	22.212053	21.417522
5	21.380534	22.458810	21.487717
6	20.965048	22.239374	21.091719
7	**18.405186**	**19.875562**	**18.551344**

Table 4.13 below regresses each variable in the model against consumer sentiment index. As the addition of further variables and lags could result in

[2]It was not possible to increase the lag due to lack of observations, however, it is believed if additional observations were available it is believed the optimal lag would higher.

alternative results, the values below should be taken only as a possible indication for further analysis and not as statistical fact.

Table 4.13: Model 1 VAR analysis

Model	1a	1b	1c	1d
const	689.32	-1158.18	-117.87	-60.16
$\text{ConsumerSentimentIndex}_{t-1}$	6.65***	9.07**	0.06	-3.12
$\text{ConsumerSentimentIndex}_{t-2}$	0.62	-6.1	1.52	1.17
$\text{ConsumerSentimentIndex}_{t-3}$	-0.50	4.99	-6.54	3.66
$\text{ConsumerSentimentIndex}_{t-4}$	2.53	-2.42	3.78	8.40
$\text{ConsumerSentimentIndex}_{t-5}$	-3.33	4.20	0.43	0.73
$\text{ConsumerSentimentIndex}_{t-6}$	2.22	7.50	6.67	-1.03
$\text{ConsumerSentimentIndex}_{t-7}$	-4.82	8.43*	0.37	-2.13
Crime_{t-1}	-	-0.08**	-	-
Crime_{t-2}	-	0.04*	-	-
Crime_{t-3}	-	-0.07**	-	-
Crime_{t-4}	-	0.06*	-	-
Crime_{t-5}	-	-0.08*	-	-
Crime_{t-6}	-	0.00	-	-
Crime_{t-7}	-	-0.13**	-	-
Negative_{t-1}	-	-	0.00	-
Negative_{t-2}	-	-	0.00	-
Negative_{t-3}	-	-	0.00	-
Negative_{t-4}	-	-	0.00	-
Negative_{t-5}	-	-	0.00	-
Negative_{t-6}	-	-	0.00	-
Negative_{t-7}	-	-	0.00	-
GardaPeople_{t-1}	-	-	-	0.02*
GardaPeople_{t-2}	-	-	-	0.03
GardaPeople_{t-3}	-	-	-	0.01
GardaPeople_{t-4}	-	-	-	0.01
GardaPeople_{t-5}	-	-	-	-0.01
GardaPeople_{t-6}	-	-	-	0.00
GardaPeople_{t-7}	-	-	-	0.01
Adjusted R^2	0.269	0.888	0.020	0.810
P-value(F)	0.0049	0.0036	0.0888	0.0132
Durbin-Watson	2.41	2.69	3.01	2.88

Note:The significance for each coefficient is given at 99% (***), 95% (**) and 90% (*). All values are multiplied by 10

Table 4.14 performs a Granger-Causality analysis on all variables to determine the direction of causality. In this case it was identified that there is no causality present between any of the variables and therefore none of the results from the VAR analysis are statistically significant.

Table 4.14: Model 1 Granger-Causality Analysis

Hypothesis	P-Value
CSI Does Not Granger-Cause GardaPeople	0.6815
GardaPeople Does Not Granger-Cause CSI	0.1882
Conclusion	No causal relationship
CSI Does Not Granger-Cause Crime	0.6280
Crime Does Not Granger-Cause CSI	0.1146
Conclusion	No causal relationship
CSI Does Not Granger-Cause Negative	0.6730
Negative Does Not Granger-Cause CSI	0.7101
Conclusion	No causal relationship

4.3.2 Garda People

Firstly all variables were tested for stationarity through the use of the Augmented Dickey-Fuller (ADF) test. Using the critical values as outlined previously (Chapter 3) there were no unit-roots identified in each variable and therefore all variables were confirmed as being stationary.

Subsequently, the lag length (p) was determined determined through the combinatory use of Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Hannan-Quinn criterion (HQC). The results of this analysis can be seen in table 4.15 were lag 4 was collectively selected as being the ideal lag.

Table 4.15: Model 2 Lag Determination

lags	AIC	BIC	HQC
1	12.202381	12.316738	12.246526
2	12.197395	12.318105	12.243993
3	12.186120	12.313184	12.235171
4	**12.169626**	**12.303043**	**12.221129**
5	12.170964	12.310733	12.224919
6	12.173313	12.319436	12.229721
7	12.176039	12.328516	12.234900
8	12.176381	12.335211	12.237694

Table 4.16 provides a summary of the results of the VAR analysis. Beginning with citations of Garda People alone, variables were added in a step-wise fashion to test if the added variable improved the fit of the model to the data. Using the Durbin-Watson value all models showed no indications of auto-correlation. The F-test indicates that all models fit the data better than an intercept-only model.

Table 4.16: Model 2 VAR analysis

Model	2a	2b	2c	2d	2e
const	56.68***	134.16***	137.55***	150.29***	147.12***
GardaPeople_{t-1}	0.28***	0.54***	0.51***	0.52***	0.51***
GardaPeople_{t-2}	0.02	0.06	0.06	0.06	0.06
GardaPeople_{t-3}	0.13***	0.10*	0.10*	0.12**	0.13**
GardaPeople_{t-4}	0.14***	0.18**	0.18**	0.16**	0.21***
Negative_{t-1}	-	-0.08***	-0.10***	-0.09***	-0.10***
Negative_{t-2}	-	-0.02***	0.00	0.00	0.00
Negative_{t-3}	-	0.00	0.00	0.02	0.02
Negative_{t-4}	-	0.00	0.00	0.00	0.00
Crime_{t-1}	-	-	0.40**	0.42**	0.43**
Crime_{t-2}	-	-	-0.37**	-0.40**	-0.39**
Crime_{t-3}	-	-	-0.25	-0.16	-0.12
Crime_{t-4}	-	-	-0.16	-0.14	-0.09
$\text{GardaStation}_{t-1}$	-	-	-	-0.82	-0.84
$\text{GardaStation}_{t-2}$	-	-	-	-0.20	-0.17
$\text{GardaStation}_{t-3}$	-	-	-	-1.93**	-1.82**
$\text{GardaStation}_{t-4}$	-	-	-	-0.05	0.15
Judge_{t-1}	-	-	-	-	0.58
Judge_{t-2}	-	-	-	-	-0.18
Judge_{t-3}	-	-	-	-	-0.45
Judge_{t-4}	-	-	-	-	-2.11**
Adjusted R^2	0.158	0.260	0.269	0.279	0.283
P-value(F)	0.0001	0.0001	0.0001	0.0001	0.0001
Durbin-Watson	2.00	2.00	2.00	2.00	2.01

Note: The significance for each coefficient is given at 99% (***), 95% (**) and 90% (*).

Table 4.17 performed Granger-causality tests on all the variables and confirms the existence of a causal between all variables and the dependent variable. It is discovered that citations of Garda people has a positive relationship with lagged values of itself at 1, 3, and 4 days. Negative sentiment in the newspapers was discovered to have a minor negative effect at 1-day lag. Hostile terms, a subset of negative terms, was also tested in place of negative. It was found to have identical relationships between the variables, however, it had a lower adjusted R^2. For this reason, negative sentiment was used for the remainder of the models. Crime is seen to be statistically significant at 95% level of significance at lag 1 and 2, however, the coefficients of 0.43

and -0.39 respectively largely cancel each other out. The addition of Garda Stations and Judges further improved the fit of the model to the data. Garda Stations at 3-day lag had a significant negative effect with a value of -1.82 and Judge mentions at a 4-day lag had a further significant negative impact with a value -2.11. A possible reason for this is because garda stations or judges may be mentioned when the investigation into the crime is finished and the punishment is being given to the offender, however, this is purely circumstantial.

Table 4.17: Model 2 Granger-Causality Analysis

Hypothesis	P-Value
GardaPeople Does Not Granger-Cause GardaStation	0.2407
GardaStation Does Not Granger-Cause GardaPeople	0.0001
Conclusion	GardaStation ⇒ GardaPeople
GardaPeople Does Not Granger-Cause Judge	0.5243
Judge Does Not Granger-Cause GardaPeople	0.0032
Conclusion	Judge ⇒ GardaPeople
GardaPeople Does Not Granger-Cause Negative	0.0001
Negative Does Not Granger-Cause GardaPeople	0.0001
Conclusion	bidirectional causality
GardaPeople Does Not Granger-Cause Negative	0.9698
Negative Does Not Granger-Cause GardaPeople	0.0001
Conclusion	Crime ⇒ GardaPeople

4.3.3 Negative Sentiment

Firstly all variables were tested for stationarity through the use of the Augmented Dickey-Fuller (ADF) test. Using the critical values as outlined previously (Chapter 3) there were no unit-roots identified in each variable and therefore all variables were confirmed as being stationary.

Subsequently, the lag length (p) was determined determined through the combinatory use of Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Hannan-Quinn criterion (HQC). The results of this analysis can be seen in table 4.18 AIC selected lag 8 while BIC and HQC selected 7. Upon testing both it was determines that lag 8 provided superior results.

Table 4.18: Model 3 Lag Determination

lags	AIC	BIC	HQC
1	36.707460	36.784198	36.737095
2	36.669724	36.804015	36.721585
3	36.648584	36.840428	36.722670
4	36.637751	36.887148	36.734063
5	36.630849	36.937800	36.749387
6	36.557305	36.921809	36.698069
7	36.161326	**36.583383**	**36.324316**
8	**36.147845**	36.627455	36.333060
9	36.165846	36.703009	36.373287

Table 4.19 provides a summary of the results of the VAR analysis. Beginning with frequency of negative terms alone, variables were added in a step-wise fashion to test if the added variable improved the fit of the model to the data. Using the Durbin-Watson value all models showed no indications of auto-correlation. The F-test indicates that all models fit the data better than an intercept-only model. GardaStations was excluded from model 3e as the Granger-Causality tests in table 4.20 identified it as not having a causal relationship with negative terms.

Table 4.19: Model 3 VAR analysis

Model	3a	3b	3c	3d	3e
const	486.21***	572.04***	608.48***	690.61***	725.18***
Negative_{t-1}	0.04	-0.05	-0.07	-0.07	-0.06
Negative_{t-2}	-0.04	-0.09***	-0.05	-0.04	-0.04
Negative_{t-3}	0.03	-0.00	0.02	0.04	0.57
Negative_{t-4}	0.00	-0.06	-0.06	-0.08	-0.06
Negative_{t-5}	0.00	-0.03	0.12**	0.14**	0.15**
Negative_{t-6}	-0.03	-0.18	0.01	0.02	0.01
Negative_{t-7}	0.59***	0.63***	0.61***	0.58***	0.61***
Negative_{t-8}	-0.06	-0.09**	-0.22***	-0.20***	-0.22***
GardaPeople_{t-1}	-	0.45**	0.43*	0.43*	0.40*
GardaPeople_{t-2}	-	0.32	0.23	0.29	0.20
GardaPeople_{t-3}	-	0.15	0.09	0.09	0.07
GardaPeople_{t-4}	-	0.52*	0.54**	0.54**	0.50**
GardaPeople_{t-5}	-	0.25	0.38	0.31	0.47*
GardaPeople_{t-6}	-	-0.15	-0.10	-0.04	-0.09
GardaPeople_{t-7}	-	-0.53	-0.59	-0.73**	-0.58
GardaPeople_{t-8}	-	0.28	0.18	0.26	0.07
Crime_{t-1}	-	-	0.28	0.47	0.06
Crime_{t-2}	-	-	-0.40	-0.64	-0.64
Crime_{t-3}	-	-	-0.19	0.19	-0.26
Crime_{t-4}	-	-	-0.00	0.03	0.03
Crime_{t-5}	-	-	-3.04***	-3.15***	-2.67***
Crime_{t-6}	-	-	-1.00	-0.58	-0.90
Crime_{t-7}	-	-	-0.01	-0.48	-0.11
Crime_{t-8}	-	-	2.69***	2.87***	2.39***
$\text{GardaStation}_{t-1}$	-	-	-	-1.71	-
$\text{GardaStation}_{t-2}$	-	-	-	-2.38	-
$\text{GardaStation}_{t-3}$	-	-	-	-6.20*	-
$\text{GardaStation}_{t-4}$	-	-	-	1.38	-
$\text{GardaStation}_{t-5}$	-	-	-	-1.17	-
$\text{GardaStation}_{t-6}$	-	-	-	-5.50**	-
$\text{GardaStation}_{t-7}$	-	-	-	7.02*	-
$\text{GardaStation}_{t-8}$	-	-	-	-4.97*	-
Courts_{t-1}	-	-	-	-	-0.51
Courts_{t-2}	-	-	-	-	-0.35
Courts_{t-3}	-	-	-	-	-1.11
Courts_{t-4}	-	-	-	-	0.24
Courts_{t-5}	-	-	-	-	-4.71***
Courts_{t-6}	-	-	-	-	-1.19
Courts_{t-7}	-	-	-	-	-1.09
Courts_{t-8}	-	-	-	-	1.80
Adjusted R^2	0.347	0.367	0.382	0.388	0.387
P-value(F)	0.0001	0.0001	0.0001	0.0001	0.0001
Durbin-Watson	2.00	2.00	2.00	2.01	2.01

Note: The significance for each coefficient is given at 99% (***), 95% (**) and 90% (*).

Negative sentiment can be seen to have a relationship with lagged values of itself at lag 5, 7, and 8. Negative sentiment at 1 week prior was the most significant at 0.61. Negative sentiment at lag 4 and 8 was 0.15 and -0.22 respectively which mostly cancel each other out. Garda People citations contained three statistically significant values at lag 1, 4, and 5 with values of 0.40, 0.50, and 0.47 respectively. From these values, it can be seen that citations of garda people on these days prior leads to an increase in negative terms today. Mentions of crime has two statistically significant mentions with large coefficients, however, these large value largely cancel each other out leaving a minor negative effect on the dependent variable. The addition of this variable into the model improved the fit of the model to the data so it was kept in regardless of the minor effect it had. Finally, the addition of Court mentions contained one statistically significant value at lag 5 with a value of -4.71. From the VAR analysis it can be seen mentions of the Courts have the most significant effect in reducing negative sentiment in the news. The addition of each variable resulted in an improved adjusted R^2 with the adjusted R^2 value of the final model being 0.393.

Table 4.20: Model 3 Granger-Causality Analysis

Hypothesis	P-Value
Negative Does Not Granger-Cause GardaPeople	0.0001
GardaPeople Does Not Granger-Cause Negative	0.0002
Conclusion	bidirectional causality
Negative Does Not Granger-Cause GardaStation	0.0001
GardaStation Does Not Granger-Cause Negative	0.1544
Conclusion	Negative $\Rightarrow$ GardaStation
Negative Does Not Granger-Cause Crime	0.0001
Crime Does Not Granger-Cause Negative	0.0015
Conclusion	bidirectional causality
Negative Does Not Granger-Cause Court	0.0001
Court Does Not Granger-Cause Negative	0.064
Conclusion	bidirectional causality

Chapter 5

Future work & Conclusion

5.1 Summary of Results

This thesis fits alongside a growing body of research which investigates sentiment analysis, the behaviour of the media and the impact of this behaviour on individuals and organisations. The thesis was broadly separated into three distinct categories, with the output from each being discussed below.

With regards to data acquisition, a high-quality corpus of 41,779 articles was collected from 23 sources, representing a mixture of national and regional publications as well as tabloid and broadsheet publications. As the selection criteria of articles required an article to mention "garda" or "police", the corpus can be considered highly relevant to An Garda Síochána. This diverse mixture of publications coupled with the number of articles collected over the two years makes this corpus a valuable source of analysis for this thesis and further research papers.

In-depth research was performed on the various different content analysis approaches which could be used for this thesis. It was identified based on the current research, that a lexicon-based approach to sentiment analysis would result in high-quality results without the need to train machine-learning models. The general Inquirer dictionary was chosen as the base dictionary based on the high quality of results that were seen from other research papers. On top of this, six additional dictionaries were researched and created which represented various aspects of the operations of An Garda Síochána. These dictionaries include mentions of crime, Garda people, Garda stations, Garda departments, courts, and their respective judges. These dictionaries were based on information from authoritative sources and created through the use of AntConc, a corpus linguistics research tool used to identify how terms are commonly presented in a given corpus.

The corpus and dictionaries were then analysed in the content affect analysis system Rocksteady. This enabled me to analyse the data based on absolute term frequency or z-scores and aggregate the data based on the source or a given time-series. Rocksteady had the ability to export this aggregated data to CSV file which enabled me to further analyse it in additional statistical programs such as GRETL, R, EViews and Excel.

The statistical analysis revealed many interesting relationships between the data. It can be seen that the Irish newspapers which form the corpus tend to use active rather than passive terms and strong rather than weak terms when discussing An Garda Síochána. It was also identified that the news media tend to use more positive than negative language when discussing An Garda Síochána, despite discussing serious crimes such as homicide, sexual offences and explosives the most frequently. This thesis further validates a statement which was said 20 years ago by (O'Connell 1999) who stated: "Typical crimes in the Irish press appear rarely in the official crime statistics and typical crimes in the official figures appear rarely in the Irish press account of crime". It was identified that the relative proportion of category 1 (Murder, Manslaughter) mentioned in the news media was 684 times greater than its relative proportion of crimes reported, thus identifying the media bias for sensationalist crimes. It was identified that general gardaí, the commissioner and Sergeants were the three most commonly cited personnel relating to An Garda Síochána. A distinction was also seen where broadsheet sources more often mentioned the higher tier position in An Garda Síochána, while the smaller regional sources dominated the lower tier titles such as inspector, sergeant and garda.

The news media was seen to cite the District Court and its judges considerably more than any other court. This was assumed to occur as regional newspapers would tend to cover all cases that occur in their district, most of which occur in the district court while broadsheet sources cover only the more serious crimes which happen in the other courts. However, this was shown to not be the case and there was no clear distinction visible between the tabloid/broadsheet publications and the courts they cited. The Special Crime Operations department was the most cited department with regional sources showing a preference for mentioning it. The mentions of Garda stations shows it could be a reliable proxy for the regional distribution of crime where all but one of the regions differed by no less than 3%.

Finally, the VAR analysis uncovered some underlying relationships between the variables. The Granger-Causality tests showed that there was a causal relationship between the independent variables in models 2 & 3 while no

causal relationship existed in model 1. It was discovered that many of the variables relating to An Garda Síochána were statistically significant in influencing the level of negative sentiment in the newspapers. Negative sentiment 5 and 7 days prior is seen to increase negative sentiment today. Garda People and crime citations are seen to have several relationships across various lags, however, these effects largely cancel each other out. Mentions of the courts at a 5-day lag is seen to significantly reduce negative sentiment in the news media today. Negative sentiment is notoriously hard to predict, however, this report managed to achieve an adjusted R^2 of 0.387 which I consider a reasonable fit. The Garda People VAR analysis also identified relationships between the variables, albeit weaker. Mentions of Garda Stations and mentions of judges were the most significant variables who both lead to a reduction in the citations of Garda people. Crime which occurred yesterday was seen to increase garda people citation today, however, this was largely eliminated by crime mentions two days prior. Finally, It was discovered that based on the data gathered over the two year period, none of the variables had any causal relationship to the Consumer Sentiment Index.

5.2 Future Work

While this thesis discovered some impressive results, there are numerous alternative approaches which could be taken which could potentially result in superior outcomes.

The most straightforward extension to this thesis would be the addition of additional data. The same process could be used, but additional data added to the corpus by either (A) selecting a longer time-frame by going further back in time or (B) selecting additional publications and adding their articles from the same period to the existing corpus. Going back further in time would enable more accurate autoregression analysis to be performed and would provide a clearer picture as to how the sentiment in the Irish news media fluctuated throughout the years. Selecting additional publications may make to corpus more representative of the entire country as it will represent more opinions; however, it also comes with some potential risks which should be considered. As this project makes use of practically all of the significant publications in Ireland, it is likely that the selection of additional publications would involve potentially "inferior quality" sources such as tabloids or online sources. These sources as identified by Kerrigan & Shaw (1985), Dahlgren & Sparks (1992) have exceptionally high levels of sensationalism which may

result is substantially higher mentions of crime and negativity. The addition of more data would also provide the opportunity of testing model 1 with additional lags and variables which could yield exciting results.

As this project used articles from newspapers, the content is well-formed and lacking slang. This ensured that it was not necessary to pre-process the data and ensured that all words in the corpus were grammatically correct. However, while these words may match terms in the dictionary, there is no guarantee that the sentiment orientation in the dictionary is correct. While general news may not require a specialist dictionary, this project focused specifically on articles relating to An Garda Síochána which could be considered a specialist domain. Therefore, it is possible that there are many terms which are not annotated with the correct sentiment orientation. For example, the term 'force' in 'police force' is considered a negative term, however, in this context that is not the case. As identified by Grimmer & Stewart (2013) using a specialist dictionary can result in vastly different results in a specialist domain when compared against a general dictionary. This thesis did attempt to make use of a police specialist dictionary, however, there was no such dictionary available online and the creation of such a dictionary was outside the scope of this project.

This thesis made use of the lexicon-based approach of sentiment analysis, specifically the dictionary-based approach through the use of the General Inquirer dictionary. As stated in Chapter 2, this dictionary is a unigram dictionary and therefore is unable to take into account contextual valence shifters such as 'not' or 'very' effectively. A possible way to extend this thesis is through the use of a prepossessing algorithm such as that used by Turney (2002) which was used to account for negation in a sentence. A further extension could be to make use of the Sentiment Orientation Calculator (SO-Cal) developed by Taboada et al. (2011). This was a specially made tool which is built on top of the General Inquirer dictionary but specifically accounts for contextual valence shifters present in the text. A further alternative approach could be to make use of Machine Learning models which have been shown in some papers to yield superior results to a lexicon-based approach alone (Pang et al. 2008). Multiple different models such as RNN, SVM and Naive Bayes could be tested and the best performing model compared against the results found in this thesis. Finally, a hybrid approach could also yield positive results as researched by Barry (2017) who discovered that incorporating a Long Short Terms Memory model along with the bag-of-words model yielded superior results across a range of metrics than a bag-of-words model alone.

5.3 Final Remarks

While this research carried out on this thesis proved the be a success, there were some challenges faced which are worth mentioning. Firstly, in creating the specialist dictionaries, there was a fine line between collecting as many distinct mentions as possible and double-counting mentions. Upon manual evaluation most dictionaries performed quite well; however, it was noticed that the Garda People dictionary had some issues. It was found that when news publications used only titles it was difficult to identify which rank the Garda was. While it was feasible to assume Mr.Harris implies the commissioner, for lower rank Gardaí and Sergeants it was found that there are many common names across the categories. In the absence of rank before the name, there was no choice but to ignore the name for fear of incorrectly labelling the individual's position. Furthermore, the Garda rank was exceptionally difficult as depending and the context, use of the term 'Garda' could refer to the organisation or an Individual.

Furthermore, at the beginning of the thesis, the Consumer Sentiment Index was identified as a key variable for research. However, oversight into the limited amount of data points meant that all variables were unable to be tested without getting spurious results. If I were to do this thesis again, I would ensure to gather more data, possibly four years, which would enable me to perform comprehensive analysis on both a daily but also monthly timescale.

Despite these challenges, I believe that the benefits of this thesis far outweight these drawbacks. Working on this project enabled me to expand my knowledge of many different areas in which I previously had little knowledge. I have learned a great deal about content analysis, in particular, sentiment analysis and the different machine-learning and lexicon-based approaches which can be employed to perform various types of content analysis. My knowledge in the area of statistical analysis has improved by order of magnitude. This project provided me with the opportunity to analyse the data using various forms of descriptive and inferential statistics. I have gained invaluable knowledge in correlation and regression analysis, particularly vector autoregression, which I have no doubt will be useful in any future research projects I carry out.

I have been exposed to a vast array of different tools which I now feel competent in using. These range from the use of LexisNexis for data collection, to Rocksteady for content analysis to R Studio, GRETL, EViews and Mi-

crosoft Excel for performing statistical analysis. The thesis has also provided me with the opportunity to develop many soft skills. Due to the long-term nature of this thesis, I was provided with the opportunity to enhance my time-management skills greatly. From the beginning I was required to distribute my workload effectively, ensuring I was making sufficient progress in my report while not neglecting other modules.

Overall I believe that this project has been a great success and I now feel confident in my skills to identify, research, implement, and analyse a research topic. I feel I have gained invaluable knowledge in the area of computer science and statistics and feel I have made a thesis which I am proud to stand behind.

Bibliography

ABC (2018), 'Irish abc newspaper circulation january to june 2018'. Online; accessed 11 March 2019.
URL: *http://www.ilevel.ie/media-blog/print/irish-abc-newspaper-circulation-january-june-2018/*

Ahmad, K. (2008), 'Being in text and text in being: Notes on representative texts', *Incorporating corpora. Clevedon: Multilingual Matters* pp. 60–91.

Ahmad, K. (2011), 'Affective computing: Sentiment, metaphor and terminology'.

Ahmad, K. (2014), 'Method and systems for calculating affect in one or more documents'. US Patent Number: 14/214,080.

Ahmad, K., Daly, N. & Liston, V. (2011), What is new? news media, general elections, sentiment, and named entities, *in* 'Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2011)', pp. 80–88.

Ahmad, K., Han, J., Hutson, E., Kearney, C. & Liu, S. (2016), 'Media-expressed negative tone and firm-level stock returns', *Journal of Corporate Finance* **37**, 152–172.

Akaike, H. (1974), A new look at the statistical model identification, *in* 'Selected Papers of Hirotugu Akaike', Springer, pp. 215–222.

An Garda Síochána (2017), '2017 annual report'. Online; accessed 21 February 2019.
URL: *https://www.garda.ie/en/about-us/our-departments/office-of-corporate-communications/news-media/garda-annual-report-2017.pdf*

An Garda Síochána (2018*a*), 'Organisational structure'. Online; accessed 21 February 2019.
URL: *https://www.garda.ie/en/about-us/organisational-structure/*

An Garda Síochána (2018*b*), 'Public attitudes survey: Q4 2018'. Online; accessed 11 March 2019.

URL: *https://www.garda.ie/en/about-us/our-departments/office-of-corporate-communications/news-media/public-attitudes-survey-bulletin-q4-2018.pdf*

Andreevskaia, A. & Bergler, S. (2008), 'When specialists and generalists work together: Overcoming domain dependence in sentiment tagging', *Proceedings of ACL-08: HLT* pp. 290–298.

Asghar, Z. & Abid, I. (2007), 'Performance of lag length selection criteria in three different situations'.

Aue, A. & Gamon, M. (2005), Customizing sentiment classifiers to new domains: A case study, *in* 'Proceedings of recent advances in natural language processing (RANLP)', Vol. 1, Citeseer, pp. 2–1.

Baiocchi, G. & Distaso, W. (2003), 'Gretl: Econometric software for the gnu generation', *Journal of applied econometrics* **18**(1), 105–110.

Barreto, H. & Howland, F. (2005), *Introductory econometrics: using Monte Carlo simulation with Microsoft excel*, Cambridge University Press.

Barry, J. (2017), Sentiment analysis of online reviews using bag-of-words and lstm approaches., *in* 'AICS', pp. 272–274.

Black, L. (2015), 'Media, public attitudes and crime', *Healy, D., Hamilton, C., Daly, Y. and Butler, M. (eds.). The Routledge Handbook of Irish Criminology* .

Brants, T. (2000), Tnt: a statistical part-of-speech tagger, *in* 'Proceedings of the sixth conference on Applied natural language processing', Association for Computational Linguistics, pp. 224–231.

Braun, P. A. & Mittnik, S. (1993), 'Misspecifications in vector autoregressions and their effects on impulse responses and variance decompositions', *Journal of Econometrics* **59**(3), 319–341.

Brill, E. (1992), A simple rule-based part of speech tagger, *in* 'Proceedings of the third conference on Applied natural language processing', Association for Computational Linguistics, pp. 152–155.

Carroll, C. E. & McCombs, M. (2003), 'Agenda-setting effects of business news on the public's images and opinions about major corporations', *Corporate reputation review* **6**(1), 36–46.

CitizensInformation.ie (2019), 'Children court'. Online; accessed 21 February 2019.
URL: *http://www.citizensinformation.ie/en/justice/courts_system/children_court.html*

Claeskens, G. & Hjort, N. L. (2008), Model selection and model averaging, Technical report, Cambridge University Press.

Cohen, B. C. (2015), *Press and foreign policy*, Vol. 2321, princeton university press.

Cook, J. A., Zhao, Z. & Ahmad, K. (2016), Stylized facts of linguistic corpora: Exploring the lexical properties of affect in news, *in* 'International Conference on Intelligent Data Engineering and Automated Learning', Springer, pp. 494–502.

Cortes, C. & Vapnik, V. (1995), 'Support-vector networks', *Machine learning* **20**(3), 273–297.

Courts.ie (2019*a*), 'District court'. Online; accessed 21 February 2019.
URL: *http://www.courts.ie/courts.ie/library3.nsf/pagecurrent/131DAE1B7E9059E580257FB*

Courts.ie (2019*b*), 'The judges'. Online; accessed 21 February 2019.
URL: *http://www.courts.ie/courts.ie/library3.nsf/pagecurrent/91D731A12A4B8A0F80257FB*

Cross, R. & Butts, C. (2005), 'Blogging for votes: An examination of the interaction between weblogs and the electoral process'.

CSO (2019), 'Irish crime categoristion system (iccs)'. Online; accessed 21 February 2019.
URL: *https://www.cso.ie/en/media/csoie/methods/recordedcrime/ICCS_V2.0.pdf*

Dahlgren, P. & Sparks, C. (1992), *Journalism and popular culture*, Sage.

Dammert, L. & Malone, M. (2004), 'Fear of crime or fear of life? public insecurities in chile', *Bulletin of Latin American Research* **22**, 79 – 101.

Dave, K., Lawrence, S. & Pennock, D. M. (2003), Mining the peanut gallery: Opinion extraction and semantic classification of product reviews, *in* 'Proceedings of the 12th international conference on World Wide Web', ACM, pp. 519–528.

Devitt, A. & Ahmad, K. (2007), Sentiment polarity identification in financial news: A cohesion-based approach, *in* 'Proceedings of the 45th annual meeting of the association of computational linguistics', pp. 984–991.

Devitt, A. & Ahmad, K. (2008), Sentiment analysis and the use of extrinsic datasets in evaluation., *in* 'LREC'.

Devitt, A. & Ahmad, K. (2013), 'Is there a language of sentiment? an analysis of lexical resources for sentiment analysis', *Language resources and evaluation* **47**(2), 475–511.

Ditton, J., Chadee, D., Farrall, S., Gilchrist, E. & Bannister, J. (2004), 'From imitation to intimidation: A note on the curious and changing relationship between the media, crime and fear of crime', *British Journal of Criminology* **44**(4), 595–610.

Duffy, B., Smith, K., Terhanian, G. & Bremer, J. (2005), 'Comparing data from online and face-to-face surveys', *International Journal of Market Research* **47**(6), 615–639.

Durbin, J. & Watson, G. S. (1951), 'Testing for serial correlation in least squares regression. ii', *Biometrika* **38**(1/2), 159–177.

Dziak, J. J., Coffman, D. L., Lanza, S. T., Li, R. & Jermiin, L. S. (2019), 'Sensitivity and specificity of information criteria', *bioRxiv* p. 449751.

Eerola, T. & Vuoskoski, J. K. (2011), 'A comparison of the discrete and dimensional models of emotion in music', *Psychology of Music* **39**(1), 18–49.

Ekman, P. (1992), 'An argument for basic emotions', *Cognition & emotion* **6**(3-4), 169–200.

Elder, R. W., Shults, R. A., Sleet, D. A., Nichols, J. L., Thompson, R. S., Rajab, W., on Community Preventive Services, T. F. et al. (2004), 'Effectiveness of mass media campaigns for reducing drinking and driving and alcohol-involved crashes: a systematic review', *American journal of preventive medicine* **27**(1), 57–65.

Engle, R. F. & Granger, C. W. (1987), 'Co-integration and error correction: representation, estimation, and testing', *Econometrica: journal of the Econometric Society* pp. 251–276.

Esuli, A. & Sebastiani, F. (2006), Sentiwordnet: A publicly available lexical resource for opinion mining., *in* 'LREC', Vol. 6, Citeseer, pp. 417–422.

Feldman, R., Govindaraj, S., Livnat, J. & Segal, B. (2010), 'Management's tone change, post earnings announcement drift and accruals', *Review of Accounting Studies* **15**(4), 915–953.

Field, A. & Miles, J. (2010), *Discovering statistics using SAS:(and sex and drugs and rock'n'roll)*, Sage.

Fuller, W. (2009), *Introduction to Statistical Time Series*, Wiley Series in Probability and Statistics, Wiley.

Galtung, J. & Ruge, M. H. (1965), 'The structure of foreign news: The presentation of the congo, cuba and cyprus crises in four norwegian newspapers', *Journal of peace research* **2**(1), 64–90.

Gerbner, G., Gross, L., Morgan, M. & Signorielli, N. (1980), 'The "mainstreaming" of america: Violence profile no. 11', *Journal of communication* **30**(3), 10–29.

Gitlin, T. (2003), *The whole world is watching: Mass media in the making and unmaking of the new left*, Univ of California Press.

Graber, D. A. (1979), 'Is crime news coverage excessive?', *Journal of Communication* **29**(3), 81–92.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1460-2466.1979.tb01714.x*

Granger, C. W. (1969), 'Investigating causal relations by econometric models and cross-spectral methods', *Econometrica: Journal of the Econometric Society* pp. 424–438.

Granger, C. W. (1980), 'Testing for causality: a personal viewpoint', *Journal of Economic Dynamics and control* **2**, 329–352.

Granger, C. W., Newbold, P. & Econom, J. (2001), 'Spurious regressions in econometrics', *A Companion to Theoretical Econometrics, Blackwell, Oxford* pp. 557–561.

Grimmer, J. & Stewart, B. M. (2013), 'Text as data: The promise and pitfalls of automatic content analysis methods for political texts', *Political analysis* **21**(3), 267–297.

Gurevitch, M., Bennett, T., Curran, J. & Woollacott, J. (1982), *Culture, society and the media*, Vol. 759, Methuen London.

Hafer, R. W. & Sheehan, R. G. (1989), 'The sensitivity of var forecasts to alternative lag structures', *international Journal of Forecasting* **5**(3), 399–408.

Hannan, E. J. & Quinn, B. G. (1979), 'The determination of the order of an autoregression', *Journal of the Royal Statistical Society: Series B (Methodological)* **41**(2), 190–195.

Hatzivassiloglou, V. & McKeown, K. R. (1997), Predicting the semantic orientation of adjectives, *in* 'Proceedings of the 35th annual meeting of the association for computational linguistics and eighth conference of the european chapter of the association for computational linguistics', Association for Computational Linguistics, pp. 174–181.

Healy, D. & O'Donnell, I. (2010), 'Crime, consequences and court reports', *Irish Criminal Law Journal* **20**(1), 2–7.

Herwartz, H. & Lütkepohl, H. (2014), 'Structural vector autoregressions with markov switching: Combining conventional with statistical identification of shocks', *Journal of Econometrics* **183**(1), 104–116.

Heston, S. L. & Sinha, N. R. (2014), 'News versus sentiment: Comparing textual processing approaches for predicting stock returns', *Robert H. Smith School Research Paper* .

Hoover, K. D. (2008), 'Causality in economics and econometrics', *The New Palgrave Dictionary of Economics: Volume 1–8* pp. 719–728.

Hu, M. & Liu, B. (2004), Mining and summarizing customer reviews, *in* 'Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining', ACM, pp. 168–177.

Jewkes, Y. & Linnemann, T. (2017), *Media and Crime in the US*, Sage Publications.

Kelly, S. (2016), News, Sentiment and Financial Markets: A Computational System to Evaluate the Influence of Text Sentiment on Financial Assets, PhD thesis, Trinity College Dublin.

Kennedy, A. & Inkpen, D. (2006), 'Sentiment classification of movie reviews using contextual valence shifters', *Computational intelligence* **22**(2), 110–125.

Kerrigan, G. & Shaw, H. (1985), 'Crime hysteria', *Magill, 18th April* pp. 10–21.

Kim, Y. (2014), 'Convolutional neural networks for sentence classification', *arXiv preprint arXiv:1408.5882* .

Knowles, J. J. (1982), *Ohio citizen attitudes concerning crime and criminal justice*, Governor's Office of Criminal Justice Services Columbus, Ohio.

Koop, G. & Tole, L. (2013), 'Modeling the relationship between european carbon permits and certified emission reductions', *Journal of Empirical Finance* **24**, 166–181.

Lacy, S. (1984), 'Competition among metropolitan daily, small daily and weekly newspapers', *Journalism Quarterly* **61**(3), 640–742.

Lasswell, H. D. (1948), 'Power and personality.'.

Last, P. & Jackson, S. (1988), *The Bristol Fear and Risk of Crime Project: A Preliminary Report on Fear of Crime*, Avon and Somerset Constabulary.

Lazarsfeld, P. F., Berelson, B. & Gaudet, H. (1944), 'The people's choice.'.

Le, Q. & Mikolov, T. (2014), Distributed representations of sentences and documents, *in* 'International conference on machine learning', pp. 1188–1196.

Liew, V. K.-S. (2004), 'Which lag length selection criteria should we employ?'.

Liu, B. (2012), 'Sentiment analysis and opinion mining', *Synthesis lectures on human language technologies* **5**(1), 1–167.

Loughran, T. & McDonald, B. (2011), 'When is a liability not a liability? textual analysis, dictionaries, and 10-ks', *The Journal of Finance* **66**(1), 35–65.

Lowry, D. T., Nio, T. C. J. & Leitner, D. W. (2003), 'Setting the public fear agenda: A longitudinal analysis of network tv crime reporting, public perceptions of crime, and fbi crime statistics', *Journal of communication* **53**(1), 61–73.

Lütkepohl, H. (2005), *New introduction to multiple time series analysis*, Springer Science & Business Media.

Lütkepohl, H. & Poskitt, D. S. (1991), 'Estimating orthogonal impulse responses via vector autoregressive models', *Econometric Theory* **7**(4), 487–496.

MacKinnon, J. G. (1991), Critical values for cointegration tests, *in* 'Eds.), Long-Run Economic Relationship: Readings in Cointegration', Citeseer.

Maeroff, G. I. et al. (1998), *Imaging education: The media and schools in America*, Teachers College Press.

Maio, P. & Santa-Clara, P. (2015), 'Dividend yields, dividend growth, and return predictability in the cross section of stocks', *Journal of Financial and Quantitative Analysis* **50**(1-2), 33–60.

McCombs, M. E. & Shaw, D. L. (1972), 'The agenda-setting function of mass media', *Public opinion quarterly* **36**(2), 176–187.

McCombs, M. E. & Shaw, D. L. (1993), 'The evolution of agenda-setting research: Twenty-five years in the marketplace of ideas', *Journal of communication* **43**(2), 58–67.

McCullagh, C. (1996), *Crime in Ireland: A sociological introduction*, Vol. 996, Cork University Press Cork.

McHugh, M. L. (2012), 'Interrater reliability: the kappa statistic', *Biochemia medica: Biochemia medica* **22**(3), 276–282.

Mehrabian, A. & Russell, J. A. (1974), *An approach to environmental psychology.*, the MIT Press.

Meijer, M. M. & Kleinnijenhuis, J. (2006), 'News and corporate reputation: Empirical findings from the netherlands', *Public Relations Review* **32**(4), 341–348.

Melville, P., Gryc, W. & Lawrence, R. D. (2009), Sentiment analysis of blogs by combining lexical knowledge with text classification, *in* 'Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining', ACM, pp. 1275–1284.

Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013), 'Efficient estimation of word representations in vector space', *arXiv preprint arXiv:1301.3781* .

Miller, G. A. (1995), 'Wordnet: a lexical database for english', *Communications of the ACM* **38**(11), 39–41.

Mixon Jr, J. W. & Smith, R. J. (2006), 'Teaching undergraduate econometrics with gretl', *Journal of Applied Econometrics* **21**(7), 1103–1107.

Mogotsi, I. (2010), 'Christopher d. manning, prabhakar raghavan, and hinrich schütze: Introduction to information retrieval'.

Mosharafa, E. & Mosharafa, E. (2015), 'All you need to know about: The cultivation theory', *Global Journal of Human-Social Science Research* .

Navigli, R. (2009), 'Word sense disambiguation: A survey', *ACM computing surveys (CSUR)* **41**(2), 10.

Niedenthal, P. M. & Halberstadt, J. B. (2000), 'Emotional response as conceptual coherence', *Cognition and emotion* pp. 169–203.

O'Connell, M. (1999), 'Is irish public opinion towards crime distorted by media bias?', *European Journal of Communication* **14**(2), 191–212.

O'Connell, M., Invernizzi, F. & Fuller, R. (1998), 'Newspaper readership and the perception of crime: Testing an assumed relationship through a triangulation of methods', *Legal and Criminological Psychology* **3**(1), 29–57.

O'Mahony, D., Geary, R., McEvoy, K. & Morison, J. (2000), *Crime, Community and Locale: The Northern Ireland Communities Crime Survey*, Ashgate.

Osgood, C. E. & Tannenbaum, P. H. (1955), 'The principle of congruity in the prediction of attitude change.', *Psychological review* **62**(1), 42.

Ozcicek, O. & Douglas Mcmillin, W. (1999), 'Lag length selection in vector autoregressive models: symmetric and asymmetric lags', *Applied Economics* **31**(4), 517–524.

Page Benjamin, I. & Shapiro, R. Y. (1992), 'The rational public: Fifty years of trends in americans' policy preferences'.

Pang, B. & Lee, L. (2004), A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts, *in* 'Proceedings of the 42nd annual meeting on Association for Computational Linguistics', Association for Computational Linguistics, p. 271.

Pang, B., Lee, L. & Vaithyanathan, S. (2002), Thumbs up?: sentiment classification using machine learning techniques, *in* 'Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10', Association for Computational Linguistics, pp. 79–86.

Pang, B., Lee, L. et al. (2008), 'Opinion mining and sentiment analysis', *Foundations and Trends® in Information Retrieval* **2**(1–2), 1–135.

Parisi, N., Gottfredson, M. R., Hindelang, M. J. & Flanagan, T. J. (1979), 'Sourcebook of criminal justice statistics, 1978', *Washington, DC: US Government Printing Office* .

Pennebaker, J. W., Francis, M. E. & Booth, R. J. (2001), 'Linguistic inquiry and word count: Liwc 2001', *Mahway: Lawrence Erlbaum Associates* **71**(2001), 2001.

Pennington, J., Socher, R. & Manning, C. (2014), Glove: Global vectors for word representation, *in* 'Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)', pp. 1532–1543.

Pesaran, M. H. (2015), *Time series and panel data econometrics*, Oxford University Press.

Piantadosi, S. T. (2014), 'Zipf's word frequency law in natural language: A critical review and future directions', *Psychonomic bulletin & review* **21**(5), 1112–1130.

Polanyi, L. & Zaenen, A. (2006), Contextual valence shifters, *in* 'Computing attitude and affect in text: Theory and applications', Springer, pp. 1–10.

Porter, M. F. (1980), 'An algorithm for suffix stripping', *Program* **14**(3), 130–137.

Posner, J., Russell, J. A. & Peterson, B. S. (2005), 'The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology', *Development and psychopathology* **17**(3), 715–734.

Press Council (2019), 'Irish national newspapers'. Online; accessed 4 March 2019.
URL: *https://www.presscouncil.ie/member-publications/national-newspapers*

Ramos, M. & Guzmán, J. (2000), 'La guerra y la paz ciudadana', *Santiago: LOM Ediciones* .

Read, J. (2005), Using emoticons to reduce dependency in machine learning techniques for sentiment classification, *in* 'Proceedings of the ACL student research workshop', Association for Computational Linguistics, pp. 43–48.

Rice, D. R. & Zorn, C. (2013), 'Corpus-based dictionaries for sentiment analysis of specialized vocabularies', *Proceedings of NDATAD* pp. 98–115.

Roberts, J. (2018), *Public opinion, crime, and criminal justice*, Routledge.

Rosenblad, A. (2008), 'gret1 1.7. 3', *Journal of Statistical Software* **25**(1), 1–14.

Russell, J. A. (1994), 'Is there universal recognition of emotion from facial expression? a review of the cross-cultural studies.', *Psychological bulletin* **115**(1), 102.

Russell, J. A. & Fehr, B. (1994), 'Fuzzy concepts in a fuzzy hierarchy: Varieties of anger.', *Journal of personality and social psychology* **67**(2), 186.

Russell, J. A., Ward, L. M. & Pratt, G. (1981), 'Affective quality attributed to environments: A factor analytic study', *Environment and behavior* **13**(3), 259–288.

Schwarz, G. et al. (1978), 'Estimating the dimension of a model', *The annals of statistics* **6**(2), 461–464.

Sellin, T. (1931), 'The basis of a crime index', *Am. Inst. Crim. L. & Criminology* **22**, 335.

Sims, C. A. (1972), 'The role of approximate prior restrictions in distributed lag estimation', *Journal of the American Statistical Association* **67**(337), 169–175.

Sims, C. A. (1980*a*), 'Comparison of interwar and postwar business cycles: Monetarism reconsidered'.

Sims, C. A. (1980*b*), 'Macroeconomics and reality', *Econometrica: journal of the Econometric Society* pp. 1–48.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A. & Potts, C. (2013), Recursive deep models for semantic compositionality over a sentiment treebank, *in* 'Proceedings of the 2013 conference on empirical methods in natural language processing', pp. 1631–1642.

Stock, J. H. & Watson, M. W. (2001), 'Vector autoregressions', *Journal of Economic perspectives* **15**(4), 101–115.

Stone, P. J., Dunphy, D. C. & Smith, M. S. (1966), 'The general inquirer: A computer approach to content analysis.'.

Strapparava, C., Valitutti, A. et al. (2004), Wordnet affect: an affective extension of wordnet., *in* 'Lrec', Citeseer, p. 40.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K. & Stede, M. (2011), 'Lexicon-based methods for sentiment analysis', *Computational linguistics* **37**(2), 267–307.

Tetlock, P. C. (2007), 'Giving content to investor sentiment: The role of media in the stock market', *The Journal of finance* **62**(3), 1139–1168.

Tetlock, P. C., Saar-Tsechansky, M. & Macskassy, S. (2008), 'More than words: Quantifying language to measure firms' fundamentals', *The Journal of Finance* **63**(3), 1437–1467.

Thorndike, E. L. & Lorge, I. (1944), 'The teacher's word book of 30,000 words.'.

Tinbergen, J. J. (1939), 'Statistical testing of business cycle theories: Part i: A method and its application to investment activity'.

Tong, R. M. (2001), An operational system for detecting and tracking opinions in on-line discussion, *in* 'Working Notes of the ACM SIGIR 2001 Workshop on Operational Text Classification', Vol. 1.

Turney, P. D. (2002), Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews, *in* 'Proceedings of the 40th annual meeting on association for computational linguistics', Association for Computational Linguistics, pp. 417–424.

Wartick, S. L. (1992), 'The relationship between intense media exposure and change in corporate reputation', *Business & Society* **31**(1), 33–49.

Wasserstein, R. L., Lazar, N. A. et al. (2016), 'The asa's statement on p-values: context, process, and purpose', *The American Statistician* **70**(2), 129–133.

Watson, D. & Tellegen, A. (1985), 'Toward a consensual structure of mood.', *Psychological bulletin* **98**(2), 219.

Weber, R. P. & Namenwirth, J. (1987), 'Dynamics of culture'.

Westfall, P. H. (2014), 'Kurtosis as peakedness, 1905–2014. rip', *The American Statistician* **68**(3), 191–195.

Whissell, C. M. (1989), The dictionary of affect in language, *in* 'The measurement of emotions', Elsevier, pp. 113–131.

Wiebe, J., Wilson, T. & Cardie, C. (2005), 'Annotating expressions of opinions and emotions in language', *Language resources and evaluation* **39**(2-3), 165–210.

Williams, P. & Dickinson, J. (1993), 'Fear of crime: Read all about it? the relationship between newspaper crime reporting and fear of crime', *The British Journal of Criminology* **33**(1), 33–56.

Wilson, S. L. R. et al. (1993), *Mass media/mass culture: an introduction.*, McGraw-Hill, Inc.

Wilson, T., Wiebe, J. & Hoffmann, P. (2009), 'Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis', *Computational linguistics* **35**(3), 399–433.

Wimmer, R. D. & Dominick, J. R. (2013), *Mass media research*, Cengage learning.

Xia, R., Zong, C. & Li, S. (2011), 'Ensemble of feature sets and classification algorithms for sentiment classification', *Information Sciences* **181**(6), 1138–1152.

Yadav, V., Elchuri, H. et al. (2013), Serendio: Simple and practical lexicon based approach to sentiment analysis, *in* 'Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)', Vol. 2, pp. 543–548.

Yalta, A. T. & Yalta, A. Y. (2007), 'Gretl 1.6. 0 and its numerical accuracy'.

Yang, Y. (2005), 'Can the strengths of aic and bic be shared? a conflict between model indentification and regression estimation', *Biometrika* **92**(4), 937–950.

Yang, Y., Liu, X. et al. (1999), A re-examination of text categorization methods, *in* 'Sigir', Vol. 99, p. 99.

Zhang, H., Yu, Z., Xu, M. & Shi, Y. (2011), Feature-level sentiment analysis for chinese product reviews, *in* '2011 3rd International Conference on Computer Research and Development', Vol. 2, IEEE, pp. 135–140.

Zhao, Z. & Ahmad, K. (2015), 'A computational account of investor behaviour in chinese and us market', *Int. J. Econ. Behav. Organ* **3**(6), 78–84.

.