2018

An Exploration of Parliamentary Speeches in the Irish Parliament Using Topic Modeling

Fiona Leheny
Technological University Dublin

Recommended Citation

Leheny, Fiona (2018). *An exploration of parliamentary speeches in the Irish parliament using topic modeling.* Masters dissertation, DIT, 2018.

An Exploration of Parliamentary Speeches in the Irish Parliament using Topic Modeling

Fiona Leheny

A dissertation submitted in partial fulfilment of the requirements of
Dublin Institute of Technology for the degree of
M.Sc. in Computing (Data Analytics)

September 2017

Declaration

I certify that this dissertation which I now submit for examination for the award of MSc in Computing (Data Analytics), is entirely my own work and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

This dissertation was prepared according to the regulations for postgraduate study of the Dublin Institute of Technology and has not been submitted in whole or part for an award in any other Institute or University.

The work reported in this dissertation conforms to the principles and requirements of the Institutes guidelines for ethics in research.

Signed:

Date:

Abstract

The only resource available in the public domain which highlights parliamentary activity is parliamentary questions. Up until the last ten years, manual content analysis was carried out to classify these. More recently, machine learning techniques have been used to automatically classify and analyse these data sets. This study analyses the verbal parliamentary speeches in the Irish Parliament (known as the Dáil) over a ten-year period using unsupervised machine learning. It does so by applying a less utilised topic modeling technique, known as Non-negative Matrix Factorisation (NMF), to detect the latent themes in these speeches. A two-layer dynamic approach using NMF is applied to extract the themes raised in these speeches at a point in time and over the entire period. The findings suggest that the themes raised vary from very niche subject matter areas to more general areas and have evolved over time. The trend in the topics raised over the entire period give an indication of what the political agenda was during these Dáil terms. Furthermore, reviewing the topics at a party and individual TD level demonstrate what their political priorities are. Conversely, reviewing the topics that parties and TDs are not discussing gives an insight into the themes that they have no interest in.

Keywords: Parliamentary questions, Parliamentary speeches, Text mining, Topic modeling, Non-negative Matrix Factorisation, Clustering.

Acknowledgments

I would like to give a sincere thank you to my supervisor, Dr. Sarah Jane Delaney, for her guidance and sharing of expertise throughout this research project. Also, a special thanks to Dr. Derek Greene for providing the data set, the publishsed tools and additional guidance along the way.

Finally, I would not have been in a position to complete this Masters course without the support and patience of my husband, Colm, and my parents. Thank you all.

Contents

List of Figures

List of Tables

List of Acronyms

Dail	**D**áil **É**ireann
LDA	**L**atent **D**irichlet **A**llocation
PQ	**P**arliamentary **Q**uestion
SVM	**S**upport **V**ector **M**achine
TD	**T**eachta **D**ála
TDM	**T**erm **D**ocument **M**atrix
NMF	**N**on-negative **M**atrix **F**actorisation

Chapter 1

Introduction

One of the largest challenges facing democracy today is people losing interest in politics (Salmond, 2014). The main reason for this in Ireland is lack of confidence in any one political party since the economic crash in 2008. There are a lot of empty promises during election campaigns. After a party sets out their policies, there is no means to track whether these policies are translated into their political agenda when they are elected into government. Countries like Ireland and Great Britain have regular political debates in the format of question time which allow parliamentary members to pose questions to the government. Questions are then debated in the format of speeches. These speeches are publicly available in the format of unstructured text. The objective of this study is to utilise unsupervised machine learning techniques to extract the themes raised in parliamentary questions in Ireland. Proving successful, this could provide a systematic tool for categorising large volumes of political text into grouped themes in the Irish Parliament (Dáil Éireann) which will facilitate analysis of how these themes evolve over time. It will also provide evidence on what parliamentary parties or politicians are concerned with in Ireland.

1.1 Research Project/problem

The analysis of Parliamentary Questions (PQs) in Ireland has been manually carried out by political scientists up until recently. Delany, Sinnott, and O'Reilly (2010) performed supervised machine learning techniques which classified questions on a local-national dimension. This automated classification of the questions proved successful. Classification techniques rely on a set of labelled or pre-classified examples supplied by a subject-matter expert. The study proposed here aims to use unsupervised machine learning techniques which does not require prior labelled examples. The technique of topic modeling will be used to cluster the unlabelled textual record of the questions into common topics or themes. Greene and Cross (2015) successfully applied topic modeling techniques to the parliamentary questions in the European Parliament. This study will adopt a similar approach to Greene and Cross (2015) but applied to Irish parliamentary questions. It will explore the application of a dynamic topic model using Non-negative Matrix Factorisation (NMF) algorithm to extract topics from the parliamentary questions in the Dáil. The topics/themes generated will be assessed using an appropriate quantitative validation method. The topics will be manually labelled based on the top terms within each topic. Finally, this data will be analysed in conjunction with supplementary data on the sitting Dáils, political parties and politicians to identify and validate trends between the themes in the questions posed and the political parties that raised them. In order to confirm the validity of the topics generated, a sample of these will be reviewed against external factors to explain the patterns identified.

Exploratory research question: *”Can a two layer NMF dynamic topic model yield coherent topics in order to highlight trends in the Irish Parliamentary speeches data set?”*

1.2 Research Objectives

The primary goal of this exploratory research project is to determine if coherent topics may be extracted from the parliamentary speeches in the Dáil using machine learning techniques. Furthermore, analysis will be completed to see if there are any trends between these topics and the politicians that raised them across their associated political parties and the constituencies that they represent. In order to achieve this goal a number of objectives need to be achieved as follows:

1. Perform a comprehensive evaluation and analysis of the existing research related to topic modeling and automated techniques for analysing political datasets.
2. Source datasets and investigate the scope and limitations of this data.
3. Perform data cleansing, merging and loading of data sets.
4. Perform exploratory pre-processing of data to generate topic modeling inputs.
5. Perform topic modelling.
6. Perform parameter fitting to include the determination of the number of topics.
7. Produce final model outputs based on new parameters.
8. Identify labels on each topic (based on top terms).
9. Merge model outputs with supplementary data for trend analysis.
10. Visualise final results.

1.3 Research Methodologies

This project is exploratory research and as such, will not commence with a hypothesis but there are quantitative and empirical aspects to this study. It relies on existing datasets which have been collected from a number of sources. It is utilising a machine learning technique (NMF) already introduced by previous research, therefore, it is a form of secondary research. Nevertheless, the objective of the study is to inform further primary research in the area of using automated machine learning techniques in the political science domain.

The aim of the study is to analyse the political speeches taking place in the Irish Parliament and highlight how the matters raised in these speeches have evolved over time. This is achieved by providing a mechanism for organising and grouping these speeches at a point in time and over the complete period being analysed. The speeches are clustered based on the latent themes within the content of those speeches by applying topic modeling. The outputs of this modeling process are analysed in detail using manual and automated techniques to confirm if it successfully grouped the data and to highlight areas of interest.

The initial task involves secondary research by conducting a comprehensive review of all existing relevant literature which pertains to automated techniques used to classify political text and specifically in the area of topic modeling.

1.4 Scope and Limitations

This work is focused on the parliamentary speeches during the verbal 'question time' within the Irish Parliament over a ten-year period. It utilises a technique (NMF) which has been proven to yield coherent topics on the European parliamentary speeches. The European parliamentary speeches are concise and tend to contain a lot of technical/legal terms. From a manual review of a sample of the Irish parliamentary speeches, in contrast, these can often range from being very short and abrupt to very long sentences that provide little context to the subject matter. As a result, there was a concern at the early stages that NMF may not perform as well on this dataset. Further to this, there was a high number of erroneous characters in the content of the speeches some of which were due to the Irish language being contained in English languages speeches. Decisions had to made on how best to cleanse and filter the content of the speeches in order to improve the coherence of the topics extracted.

Finally, there are a large volume of parliamentary questions which are answered by written responses and no analysis was completed on these datasets as a decision was made to focus on the parliamentary questions which are answered verbally in the Irish Parliament.

1.5 Document Outline

The following chapters in this document are structured as follows.

Chapter 2 introduces machine learning techniques to address text mining problems under the headings of supervised and unsupervised learning. Following that, a detailed account is given of the unsupervised machine learning techniques relevant to this research which includes topic modeling and NMF. The theory and applications of the techniques which are implemented in this study are introduced here. The chapter concludes with a summary of the approaches that will be applied and the motivation behind the study.

Chapter 3 describes the methodology and design of the research undertaken and explains the rationale behind the design decisions including the related theory underpinning them. It uses the CRISP-DM framework where appropriate to structure the design commencing with an outline of the data within the political science domain, followed by an understanding of the data and the data preparation phase. This chapter also discusses how the model is built, evaluated and further analysed. It concludes with a discussion on the strengths and limitations of the proposed design approach.

Chapter 4 describes the technical implementation and the results at each stage in the process. It provides a critical evaluation of the modeling results and the final exploratory analysis carried out.

Chapter 5 provides a summary of the work carried out with an overall evaluation of the research project and concludes with suggestions for future work.

Chapter 2

Review of existing literature

This chapter provides a detailed review of the relevant literature that pertains to the subject matter areas of text mining techniques in general and more specifically in relation to unstructured text generated in the political science domain. The first section will give an overview of the political agenda within Parliaments by outlining the role that parliamentary speeches play. Following this, the theory and applications of the latest text mining techniques will be introduced making a distinction under the headings of supervised versus unsupervised machine learning algorithms. Next the text mining problem is discussed with regards to the Political Science arena, looking at the various techniques used to date and focusing in on the problem at hand. This will then lead into a detailed review of the topic modeling algorithm NMF which is chosen to address the challenges with respect to this text mining problem.

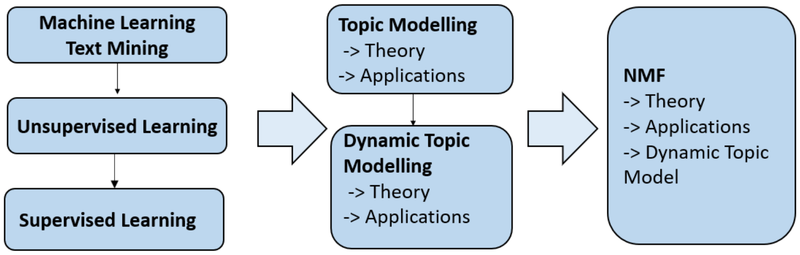


Figure 2.1: Literature Review Chapter layout

2.1 Irish Parliament Overview

Dáil Éireann, known as the Dáil for short, is the House of Representatives in the Irish Parliament and the Senate is known as Seanad Éireann. The Oireachtas is the name given to the national democratic parliament of Ireland which consists of the President, Dáil Éireann and Seanad Éireann. The people elect the members of the Dáil who are known as Deputies or TDs (Teachta Dála) who are associated with regional areas known as constituencies. The size of a constituency differs across the country but the law allows for at least one TD in the range of 20,000 to 30,000 people. A TD can exist as a member of a political party or as an independent and by law a general election is held at least every 5 years to select the members of the Dáil. The primary role of TDs and Senators is to pass laws and they also participate in debates making representations on the part of interest groups and their constituents.

2.2 Parliamentary Questions

One key process that TDs are involved in is parliamentary question time. The Dáil allocates a timeslot on Tuesdays, Wednesdays and Thursdays whereby any TD may pose a question to a Minister about matters related to the government department for which they are responsible for. This question time process is used as a tool by TDs to challenge the government making them accountable to the Dáil. The head of the Irish Government or Prime Minister, known as the Taoiseach, answers questions on Tuesdays and remaining Ministers answer on a rotating schedule. This ensures that there is opportunity to question or challenge each Minister on a regular basis. TDs must give a notice period of 3 and 5 days for written and oral questions respectively. On a particular day, only 5 questions are given priority for oral answer to ensure that adequate time is allowed to deal with them. Each question can generally result in a short dialogue by way of supplementary questions. Before a question can appear on the Dáil order record, it is reviewed to ensure that the question complies with the rules of the House. Oral questions not answered on the day may either be held over until the next time the relevant Minister answers oral questions or be addressed by means of a written reply.

As part of this question time, each opposition party leader is also allocated a short two-minute slot to pose a question to the Taoiseach about an ongoing matter of public importance known as Leader's questions. The Taoiseach's reply is cut to three minutes only and the opposition leader is also free to follow up with another question that is less than one minute. The final reply from the Taoiseach is allocated a slot of no more than another minute. All of these questions are monitored by the Ceann Comhairle or chair of Dáil Éireann who ensures that the rules of the House are abided by and TDs and Ministers are treated in a fair manner.

Question Time can be topical and sometimes heated and therefore can attract substantial media attention. It is argued by Salmond (2014) that this media coverage has the capacity to sway public opinion and political participation due to the high frequency of it over the parliamentary terms. As a result, question time is a critical tool for the political agenda and as such is fully documented on the Dáil record. This in turn generates a large volume of text all of which is published online [1] in documents and webpages. The availability of this data online has the potential to provide a valuable resource for gaining insight into not only the behaviours and attitudes of political parties and their elected representatives but also their political focus over time.

Parliamentary questions are a feature of most advanced democratic countries and is generally the only time that a government relinquishes control to the opposition (Salmond, 2014). One school of thought is that question time which involves quick-witted and impromptu debate between political leaders can enhance the general public's engagement with the political process and have higher election turnout rates (Salmond, 2014). Similar to the Irish Parliament, the question time in other countries such as Great Britain, Germany, Denmark, Canada and New Zealand take the format of an oral or written question and offers a tool for questioning both policy and constituency-oriented matters. Each of these countries may have slightly different rules and practises but, overall, it aims to provide citizens with an accessible and frequent

[1] http://www.oireachtas.ie/parliament/oireachtasbusiness/parliamentaryquestions/

resource of topical political information. Martin (2011) suggests that the questions posed can reveal the behaviour or role orientation of a politician or party. Dandoy (2011) provides evidence that the number of written questions in the Belgian Parliament, from 1995 to 2007, is explained by a number of key factors. The most obvious explanation relates to the size of the parliamentary party group and the government versus opposition status of the party. However, Dandoy (2011) argues further that the solidity, cohesiveness and discipline within a political party is mirrored in their pattern in question time. Notwithstanding this, it is worth acknowledging that the type of political debate witnessed in question time in Ireland and across other countries, based on the author's personal opinion, is often simplistic and deliberately antagonistic. Consequently, this may lead to negative effects on the citizen's engagement in the political process.

PQs span a diverse range of topics from very focused themes, such as highlighting the need for funding of a hospital or school, to very broad issues such as the state of the healthcare sector. Rozenberg and Martin (2011) analysed this usage across the British Parliament and illustrated that the non-disclosure of politically embarrassing information that subsequently is revealed by way of a PQ can have severe consequences for the reputation of a government. A prime example of this took place in the Dáil in 2017 when a PQ submitted by Deputy Alan Kelly of the Labour Party resulted in information being released which led to the resignation of the Minister of Justice, Frances Fitzgerald. Alongside this, the failure to answer a PQ accurately can also be damaging to a government, therefore, a lot of time and resources are expended within the Irish Public Sector answering PQs. Whilst the process itself can provide transparency in how public money is being spent, at the same time, a large number of resources are tied up answering these questions in state departments and agencies. Sometimes little thought is put into what they are asking by churning out the same questions at the same time each year. For this reason, the quality of questions posed should be challenged more and questions should not just be a numbers game for the TDs that raise them. Therefore, the uses to which PQs have been put to in the Dáil over a ten-year period is an important aspect to this project and will be discussed further in line with the research findings.

Another interesting aspect to the research carried out by Rozenberg and Martin (2011) related to a comparison between oral and written PQs. In the majority, oral questions tend to deal with more general policy issues whereas written questions are specific and detailed in nature (Martin, 2011). Also, the number of written questions increased dramatically over the period analysed from 2000 to 2010 across a number of democratic countries. Contrary to this, the trends in oral questions are less comparable across the same countries analysed (Rozenberg & Martin, 2011). Another key factor is the level of media attention given to oral questions of which there is notably less publicity given to written questions. This level of publicity on oral questions measures up against the fact that the questions themselves are more general and attempt to reach the wider public audience that are given to them. Written questions fulfil a different role of scrutinising the past and planned future activities of the government by way of precise and detailed questioning (Rozenberg & Martin, 2011). Oral questions are more suited for political theatrical controversies and heated debate. Notwithstanding this, both oral and written questions serve as tools for the accountability of the government in most democratic countries today (Rozenberg & Martin, 2011).

2.3 Techniques for analysing Parliamentary Questions

Manual content analysis of parliamentary questions has been carried out in Ireland (Martin, 2011), Belgium (Dandoy, 2011) and Turkey (Bulut, 2016). This research proved that the more organised and disciplined parties lead to more active question time (Dandoy, 2011). Also, unsurprisingly there are more questions from opposition parties (Dandoy, 2011) with their prime focus being on self-benefit and harming the government (Bulut, 2016). Further to this, Bulut (2016) suggests that oral questions are used to illustrate a legislators position to their constituents.

The analysis of parliamentary questions has been carried out manually by Political scientists up until recently and with the explosion of new machine learning techniques there is now a shift to automated techniques to categorise parliamentary questions. Delany et al. (2010) performed supervised machine learning techniques to classify questions in the Dáil in Ireland. This approach classified questions from 1922 to 2008 on a local-national dimension which proved automated techniques were successful. It illustrated that there was a lack of evidence within the questions analysed to prove the role clientelism plays in Irish politics. Classification techniques used in the experiment by Delany et al. (2010) rely on a set of labelled or pre-classified examples supplied by a subject-matter expert within the political science domain. The research undertaken here does not have access to labelled data or such experts so it is necessary to explore other machine learning techniques.

Topic modeling, which is an unsupervised learning technique, has been applied in the political science domain. It is useful as a tool to cluster large amounts of political text into topics. An experiment conducted by Greene and Cross (2015) successfully used topic modeling to cluster questions in the European Parliament (EP). There are many variations of topic modeling. These range from probabilistic approaches, such as hierarchical LDA and the pachinko allocation algorithm used by Blei (2012), to Non-negative matrix factorization (NMF) by Greene and Cross (2015). The latter of these will be discussed in more detail shortly. For the moment, it is enough to understand that topic modeling is a technique which may be used to extract the themes from an underlying body of text without the need for prior subject-matter knowledge. Therefore, topic modeling is one of the main research areas within this study. It will be used to cluster the unlabelled textual record of the questions into common topics or themes and provide a tool for identifying which topics a politician has selected to address thus providing insight into their political activity. In summary, this study heavily utilises the findings from the research carried out by Greene and Cross (2015) and applies this knowledge to Irish parliamentary questions.

2.4 Challenges and Limitation of PQs

It is also necessary to be mindful of the limitations that PQs pose when using them as a tool to analyse politicians and political parties. Whilst PQs can reveal the theme of questions being asked, it is a leap of faith to deduce the preferences from such behaviour (Rozenberg & Martin, 2011). It will not provide the motivation behind the elected representative's behaviour. PQs can also be used as a mechanism to alert information to ministers, electors and other interest groups, as well as retrieving information, having little or no interest in the answer. Also, the rules and procedures of question time in the House would have evolved over time so one would expect to see changes in trends as a result of this. According to Rozenberg and Martin (2011), the low-level costs and restrictions associated with submitting PQs provides a low-cost tool and opportunity for politicians to perform their role rather than having to spend hours in their constituencies witnessing the issues themselves. PQs can be seen as choices made in how they wish to perform their political duties which in itself is an interesting fact. Finally, it may not be entirely accurate to argue that they serve as an all-encompassing oversight tool. However, it is the sole repository of information that exists in the public domain that enables the analysis and interpretation of the behaviours and interests of the parliamentary members in modern countries today (Rozenberg & Martin, 2011).

2.5 Overview of Machine Learning

Machine learning is an application of artificial intelligence (AI) that enables systems to automatically learn and improve from experience without being explicitly programmed to do so. The process commences with samples of data known as observations. The main focus is on the development of computer programs that can analyse and use data to learn automatically without human intervention. The program analyses this data by looking for patterns to enable better decision-making in the future. It makes use of complex algorithms to automatically build models which represent the relationship between a set of inputs and one or more outputs. The algorithm attempts to model the relationship or behaviours within the data.

Machine learning is divided into 2 main areas of learning: supervised and unsupervised learning. Supervised learning is when prior knowledge of the data is known in the format of previously labelled examples. The process of learning uses these labelled examples to train the model to predict the outcomes for new and unknown examples. Unsupervised learning arises when there is no pre-labelled data and other characteristics in the data must be used. The machine learning approach chosen will depend on the type of data that is available. For both approaches, the data being modeled needs to be represented as a vector of features which are the characteristics that describe the observation. For structured data, these features are generally columns in a database. In the case of textual data, which is unstructured, there are no such features. Therefore, textual data needs to be converted into a representation that can be used in machine learning algorithms.

The most common and simplest way to represent textual data is using a Term Document Matrix (TDM). This representation tokenises the text into a bag of words. Each word is a feature and the representation of the textual data is the value of the feature which is related to the frequency of occurrence of the word in the document. Using this representation, the order of the words is not considered so is not suitable for natural language processing such as speech recognition where the ordering of the words is important. The dimensionality of the bag of words approach is very high and can contain a number of sparse features. Therefore, it is necessary to decide which words are the key terms whilst discarding terms that are deemed insignificant to the meaning of the text. This can be based on a number of methods such as using the most frequent words to extract the meaning. More advanced information theoretic techniques such as information gain or probability theory can be utilised. After the key terms have been identified, the next key step is to use these terms to classify/cluster the body of text. Text mining problems can be systematically addressed by machine learning techniques therefore the following subsections will elaborate on relevant text mining techniques under the headings of supervised and unsupervised machine learning.

2.5.1 Supervised Machine Learning

Supervised learning is a method of training a model to learn the relationship between a set of independent features and a dependent or target feature. The target feature can be categorical or numerical and is known as classification and regression respectively. The learning is based on pre-labelled examples used in the training process.

The classification problem is represented by a function f with f(x) = y where x represents the set of independent features and y is the outcome or target feature. The labels or categories in the target feature are agreed upfront. Each observation can relate to zero, one or multiple categories in the target. As discussed earlier, in the case of textual data, the features are extracted using the bag of words approach and represented as a TDM. The features are the terms within the text with their corresponding values representing the occurrence of the word within the text or document.

The modeling process can be split into the training and prediction phases as illustrated in figure 2.2.

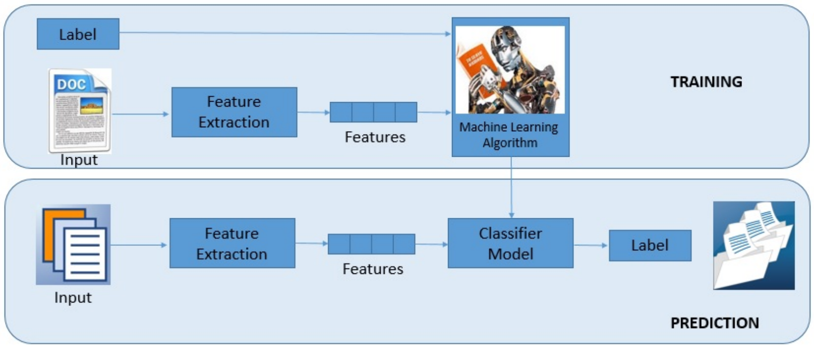


Figure 2.2: Supervised Machine Learning process

The accuracy of the model is validated on a different portion of the data known as the test dataset. From this, it is possible to deduce if the model will generalise well to the population or if it is overly aligned to the data known as overfitting. When models are overfit, this means that they can produce highly accurate results for the training dataset but predict less accurately for new datasets which is undesirable. The ability of a model to predict well for new data sets is known as generalisation.

Text classification can be implemented using a number of different supervised machine learning algorithms such as Support Vector Machines, Naive-Byes and Neural Networks but they are all based on the same principles as illustrated in figure 2.2. The algorithm chosen will depend on the text being analysed and the level of control required. To illustrate how one of these algorithms work, Support Vector Machines (SVM), discovered over fifty years ago, will be taken as an example as it is widely used in text classification problems today. In simple terms, a SVM model is a representation of the data as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New data examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall. It was originally used as a binary linear classification system with only two categories allowed (Basu, Walters, & Shepherd, 2003). For linearly separable data, this means we can draw a line on the graph which separates the two classes. In mathematical terms, a hyperplane is then employed based on the line that separates the two classes and SVM attempts to orientate the hyperplane so that it is furthest away from the nearest data points from both classes. These data points are known as the Support Vectors and the distance between them and the hyperplane is the margin (Basu et al., 2003). In the 1990's, Boser, Guyon, and Vapnik (1992) extended the use of SVM to handle data which is not linearly separable by transforming the data using a kernel function into a higher dimensional space. The process is the same in that the SVM's aim is to maximise the margin based on a small subset of training examples which generates a quadratic mathematical problem to be resolved.

A SVM may be used for classifying text by training the model on pre-labelled examples. It can handle large feature sets. It has been used in the classification of images and is widely applied in science in the classification of proteins.

This study involves text mining. The text is stored in individual webpage documents and a machine learning method is required to automatically cluster these documents in some manner for storage, retrieval and analysis purposes. Supervised and unsupervised machine learning methods are applied in text mining problems. In simple terms, whichever method is applied, the goal is to automatically assign a label or piece of metadata to each document. In this problem, the order of the words does not matter and there are no previously labelled examples available to this study. Therefore, supervised machine learning methods cannot be adopted and it is more appropriate to conduct most of the research in the area of unsupervised learning.

2.5.2 Unsupervised Machine Learning

Unsupervised machine learning can be used as a method of learning when there are no pre-labelled examples available. This is hugely beneficial with the abundance of unlabelled data now available online. It also means that there is no over reliance on subject matter experts at the initial analysis stages, however, they may be required at a later point to review the clusters or groupings to ensure they make sense in the context of the specific knowledge domain. The process allows for the identification of patterns in the dataset where there are common characteristics and provides natural groupings or clusters in the data. Unsupervised machine learning can also be used for the identification of outliers such as fraud detection. As a result, it may be used as a precursor to a supervised machine learning problem to remove potential erroneous data which will ultimately improve classification accuracy known as semi-supervised learning.

K-means clustering is the most commonly used unsupervised clustering method as it is easy to implement and can tackle large data sets. Some examples of where it is applied includes market segmentation and price segmentation. It is also used for dimensionality reduction. The original algorithm, which was introduced in a lab founded by Alexander Bell in the mid-20th century, became known as the Lloyd's algorithm. The algorithm is initialised by choosing a random set of clusters and for each record the closest centre or cluster is found by using a measure such as the euclidean distance or cosine similarity. After assigning all records, new cluster centres are assigned based on the average coordinates of the records which make up the cluster. The algorithm repeats and reassigns records based on the new cluster centres. This process continues until the algorithm has converged and the centres do not change (Wu, 2012). Before the algorithm can commence, the number of clusters (k) needs to be chosen as an input parameter and the centroids for these clusters may be chosen at random for initialisation. There are other methods for choosing the centres such as the Forgy method which disperses the initial centroids out from the centre of the data set. The main challenges with the algorithm is having to decide on the value of k as using an unsuitable value can lead to poor quality results. However, if this value is not known, there are different methods for overcoming this problem such as the elbow method, cross-validation and information criterion approaches (Wu, 2012).

2.6 Topic modeling

Topic modeling is an unsupervised machine learning technique which provides a tool for clustering textual data. It is widely used in the domain of text mining to determine the latent semantic structure or topics from a text corpus e.g. news articles, tweets, political speeches etc. These topics can be extracted from the co-occurrences of words across documents in the corpus being analysed. The goal is to identify the topics which best describe the data. As the algorithm does not require prior labelling, it enables the organisation and analysis of large volumes of data at a scale that would be infeasible by humans. The high-level approach used in topic modeling is illustrated in figure 2.3. The output of topic modeling is a collection of k clusters of documents which are similar and can be interpreted as topics.

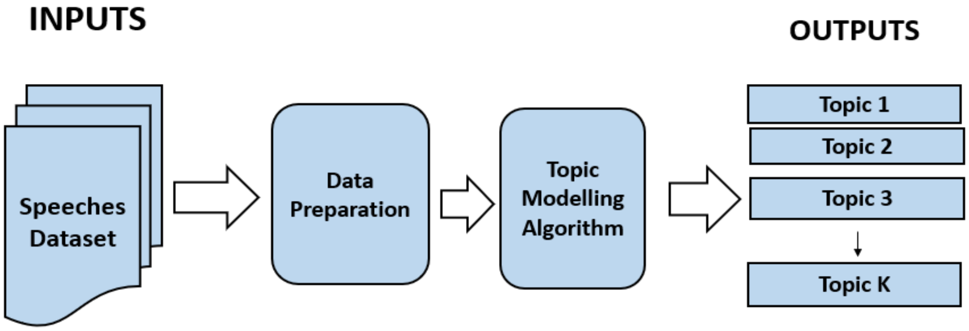


Figure 2.3: Topic Modeling Approach

2.6.1 Applications

One of the earliest topic models was invented by Papadimitriou, Tamaki, Raghavan, and Vempala (1998) but the most widely used topic model over the past 15 years, Latent Dirichlet allocation (LDA), was introduced by Blei, Ng, and Jordan (2003). This latter topic model is based on the principles of probabilistic latent semantic analysis (PLSA), which was invented by Thomas Hofmann, but goes a step further to producing a probabilistic model at the level of the documents (Blei et al., 2003). Early applications of topic modeling (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990) were used for the purpose of latent semantic indexing of documents. Since then, most of the focus has been on topic modeling which utilises probabilistic mathematical methods. There are many variations of the probabilistic LDA approach such as hierarchical LDA and the pachinko allocation (Blei, 2012).

Blei (2012) describes a topic model as a statistical tool that facilitates the analysis of unstructured text in a document to highlight their underlying themes. The model output is an organised set of documents into their topics or themes. The technique has been used in sentiment analysis (Lu, Ott, Cardie, & Tsou, 2011), social network analysis (H. Lee, Hong, & Kim, 2015), songs (Laitonjam, Padmanabhan, Pujari, & Lal, 2015) and in news text (Li, Shang, & Yan, 2016). Ahmadi, Tabandeh, and Gholampour (2016) and S. Lee, Kim, and Myaeng (2015) illustrate an enhanced combined classification and topic modeling approach to first cluster the text in a set of reduced topic features. Topic models were originally discovered in the area of text mining but have also been used in other areas such as bioinformatics and image retrieval (Blei et al., 2003). More recently other mathematical based algorithms have been introduced which attempts to find the model that produced the data such as Singular Value Decomposition (SVD) (Arora, Ge, & Moitra, 2012). The foundations of this lie within linear algebra and is based on the factorisation or decomposition of a matrix into the product of other matrices. Non-negative matrix factorization (NMF) is another type of matrix factorisation algorithm which was published by D. D. Lee and Seung (1999).

LDA and NMF have both been applied to the parliamentary speeches in the European Parliament. Further to standard methods of topic modeling, which do not consider temporal data, a dynamic topic model was developed (Blei & Lafferty, 2006). This type of model considers how topics evolve over time. More recently, Greene and Cross (2017) introduced a new two-layer approach to the standard NMF which will be discussed in further detail in the next section.

2.6.2 LDA

The LDA algorithm is based on the principles of PLSA which models each word in a document as a representation of mixture components seen as topics. Each word is generated from only one topic but there exists multiple topics. Each document can then be seen as proportions of these topics which the words relate to and in turn can be represented by a probability distribution on the set of topics. LDA goes beyond PLSA as it uses the de Finetti theorem to consider mixture models and probability distributions. It assumes the principle of exchangeability which means that the ordering of the documents, as well of as the ordering of the words within a document in a corpus, is irrelevant (Blei et al., 2003). Therefore, both the words and the documents can be represented as a probability distribution on the topics. In mathematical terms, given k topics, $P(w_i \mid z)$ is the probability distribution of the i^{th} word given topic z. $P(z_i = j)$ is the probability that the j^{th} topic was sampled for the i^{th} word and $P(w_i \mid z_i = j)$ is the probability of the i^{th} word under the j^{th} topic. The model is represented by the following equation where k is the number of topics (Steyvers & Griffiths, 2007):

$$P(w_i) = \sum_{j=1}^{k} P(w_i \mid z_i = j) P(z_i = j) \tag{2.1}$$

The above equation is further summarised in the literature using $\phi^j(w \mid z = j)P(z)$ and $\theta^d = P(z)$. LDA uses the principles of PLSA but introduces the dirichlet distribution to handle the generalising of θ to new documents which are not in the original corpus (Blei et al., 2003). The LDA algorithm uses Markov Chain Monte Carlo sampling such as the Gibbs algorithm to infer the correct model with the end goal being finding the model which can be used to reproduce the document known as a generative model. It iterates through all documents, word by word, each time estimating the topic that the word was sampled from given prior assignment of words to topics. The LDA model faces similar challenges to the k-means clustering algorithm as the number of topics has to be selected upfront.

2.7 NMF

2.7.1 Theory of NMF

As previously mentioned, there has been significant amount of research conducted on probabilistic methods of topic modeling. This approach considers a topic to be represented by a probability distribution of words. Another method of topic modeling is Non-negative Matrix Factorization (NMF). This method is also an unsupervised machine learning approach used to extract the underlying themes or topics from a corpus. NMF is a useful algorithm for reducing a large dataset into representative attributes by the reduction of the dimensionality of non-negative matrices. The document corpus is represented as a Term Document Matrix (TDM) and the algorithm reduces the data into 2 factors which do not contain any negative values. These 2 factors can then be represented by the addition of a set of two non-negative basis vectors. NMF has an inherent clustering mechanism as it generates clusters from the columns of the input matrix (i.e. the terms) which provide the basis for the topics generated. The documents where the terms are sourced can therefore be viewed as the additive combination of multiple topics which can be described as a mixed membership model. However, it is also possible to consider a single membership model or disjoint datasets or clusters where a document can relate to one topic only. This will be discussed again within the design phase of this project as to which approach will be used but at this stage it is enough to know that the outputs produced will allow for both mixed and single membership models.

In mathematical terms, NMF is based on linear algebra and is an algorithm used for matrix factorisation with the additional constraints of non-negative terms. It seeks to reduce a matrix into the dot product of two factors so that they do not contain any negative values. It can be applied for the purpose of clustering a corpus of documents. It starts by representing the corpus as a nxm document-term matrix, A. It then reduces A to W and H, a documents to topics factor and topics to terms factor respectively as highlighted in fig 2.4.

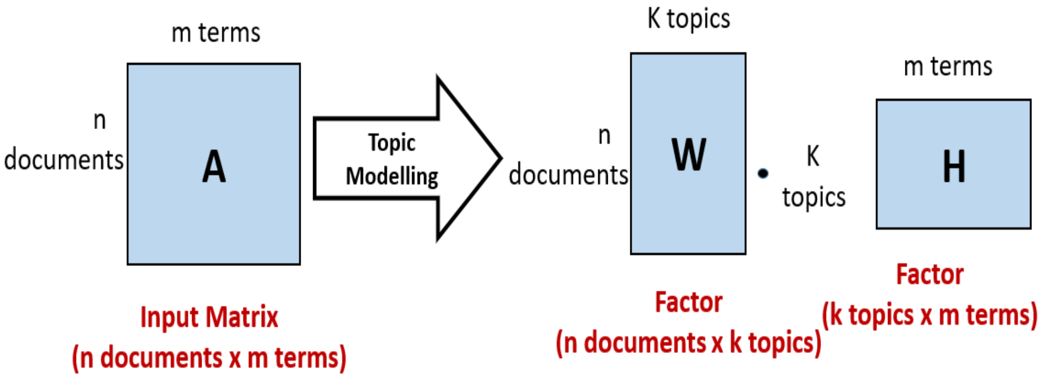


Figure 2.4: Topic modeling using NMF

Each factor can be modelled as the additive combination of a set of non-negative basis vectors which generates an inherent clustering of the data. These clusters may be viewed as topics when dealing with a text mining problem to generate a topic model. Each document within the corpus can be represented by additive combination of multiple overlapping clusters or topics.

There are different methods for finding W and H such as the multiplicative update rule used by (D. D. Lee & Seung, 1999) illustrated in figure 2.5.

$$\boxed{\begin{gathered} W_{ia} \leftarrow W_{ia} \sum_{\mu} \frac{V_{i\mu}}{(WH)_{i\mu}} H_{a\mu} \\ W_{ia} \leftarrow \frac{W_{ia}}{\sum_{j} W_{ja}} \end{gathered}} \quad \boxed{H_{a\mu} \leftarrow H_{a\mu} \sum_{i} W_{ia} \frac{V_{i\mu}}{(WH)_{i\mu}}}$$

Figure 2.5: NMF Multiplicative Update Rule.

The iteration of these update rules results in the convergence of the objective function to a local maximum preserving the constraint of non-negative values. The update rule also constrains W to sum to unity. The objective function is defined as in equation 2.2.

$$F = \sum_{i=1}^{n} \sum_{\mu=1}^{m} [V_{i\mu} \log(WH)_{i\mu} - (WH)_{i\mu}] \tag{2.2}$$

Greene and Cross (2015) uses the fast alternating least squares variant of NMF as introduced by Lin (2007) when analysing the European parliamentary data set. Alsongside this, the Non-negative Double Singular Value Decomposition (NNDSVD) approach (Boutsidis & Gallopoulos, 2008) was also applied to generate initial factors to ensure a deterministic output and better quality topics.

2.7.2 Example of NMF

Take an example of a document to term matrix with 6 documents and 9 terms which generates 2 topics. As illustrated in figure 2.6 on page 23, the shading represents the frequency of each term within each document as a TDM. Once NMF is applied, there are two output factors containing the document to topic and term to topic weightings as illustrated in figure 2.7 on page 23. The shading represents the weighting of the documents and terms to the topics/clusters generated. From this example, we can see that topic 1 relates to criminal activity based on the highest weighted terms. Document 1 and 2 have the highest weightings for this topic. Also, topic 2 could relate to deaths in healthcare with document 3 being most related to this topic.

2.7.3 Application of NMF

As previously mentioned NMF can be used in text mining to cluster similar documents together. Initially D. D. Lee and Seung (1999) used it to learn parts of faces and compared it against Principal Component Analysis (PCA) and Vector Quantization (VQ) which in contrast can only learn holistically at the whole object level. More recently, it has even been used in the investigation of cancers to discover patterns in cancer mutations (Wang, Wang, & Gao, 2013).

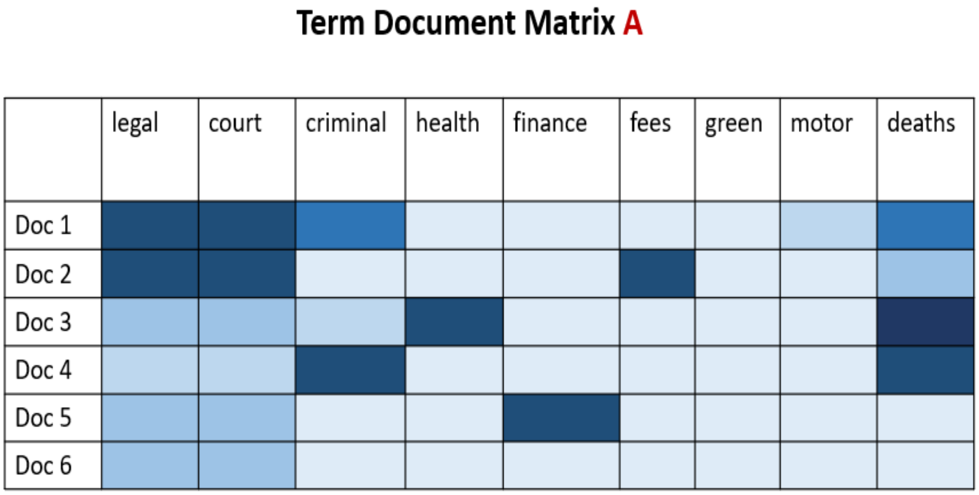


Figure 2.6: NMF Topic Model Example of TDM (A)

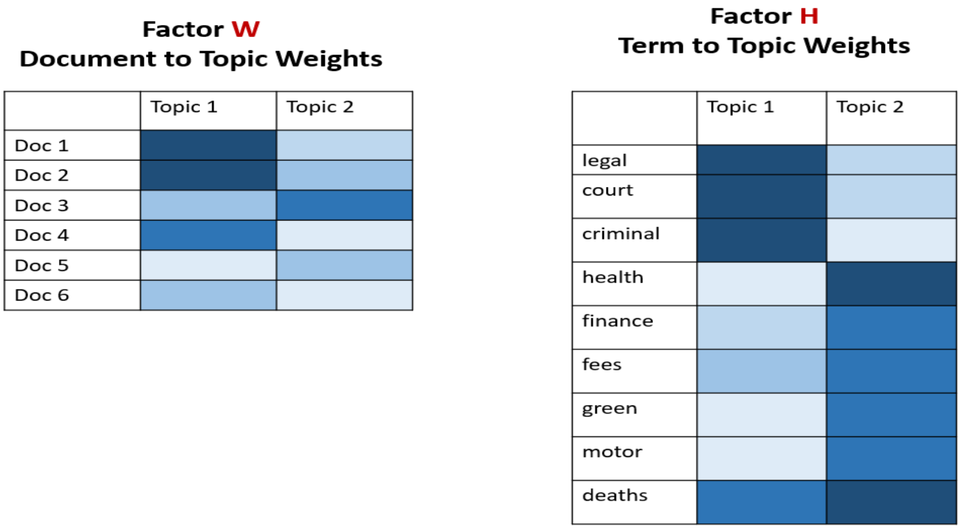


Figure 2.7: NMF Topic Model Example of W and H factors

2.7.4 Dynamic Topic modeling

Topic modeling approaches have been used in both the European parliament (Greene & Cross, 2015) and the U.S Senate (Quinn, Monroe, Colaresi, Crespin, & Radev, 2010). Both approaches demonstrated the usefulness of this approach with little need for human intervention. These approaches also utilised the idea of dynamic topic modeling which can track how language changes and the evolution of related topics over time. Standard topic modeling methods do not consider the ordering of documents and this is not suitable where documents are time-stamped. Quinn et al. (2010) utilised a combination of the Dirichlet prior algorithm and the Dynamic Linear Model (DLM) framework to capture the temporal aspect of the data. Greene and Cross (2015) introduced a new two-layer dynamic topic model by applying NMF to the time-stamped data twice. This two-layer approach is illustrated in figure 2.8.

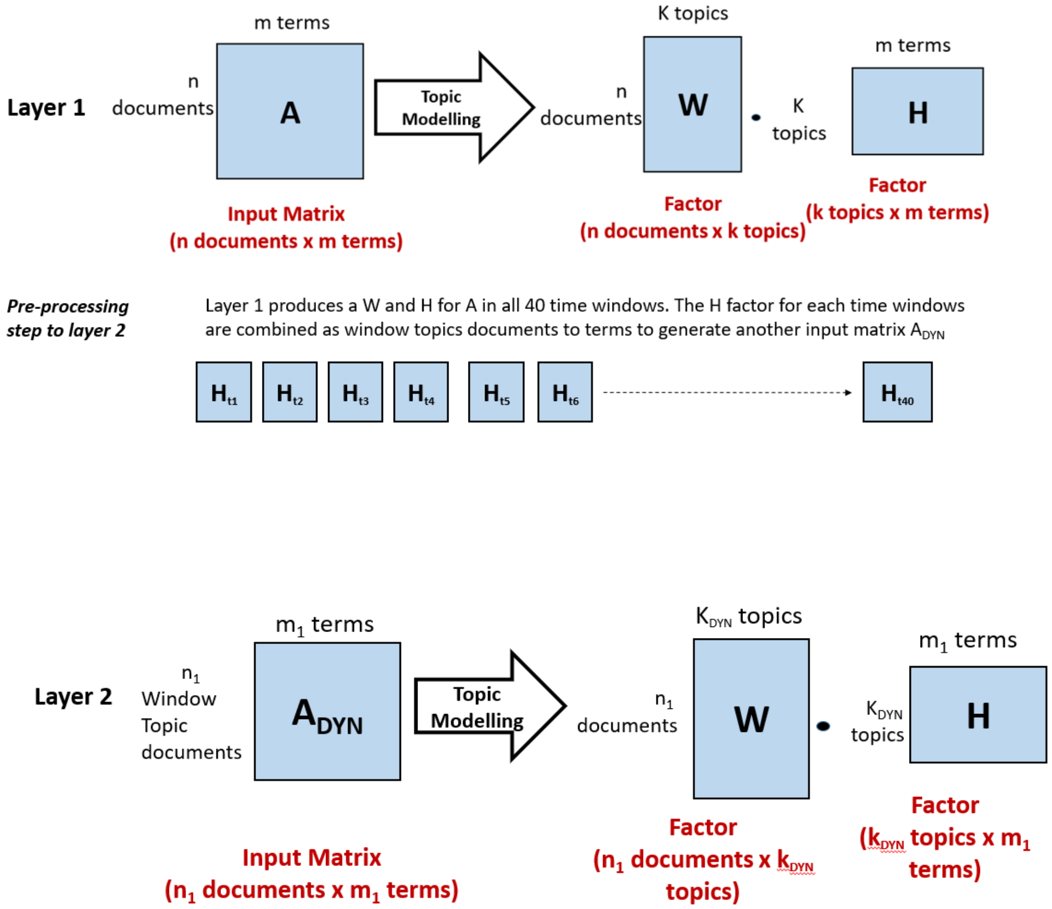

Layer 1
m terms
n documents
A
Topic Modelling
K topics
n documents
W
K topics
m terms
H
Input Matrix (n documents x m terms)
Factor (n documents x k topics)
Factor (k topics x m terms)
Pre-processing step to layer 2
Layer 1 produces a W and H for A in all 40 time windows. The H factor for each time windows are combined as window topics documents to terms to generate another input matrix A_{DYN}
H_{t1} H_{t2} H_{t3} H_{t4} H_{t5} H_{t6} H_{t40}
Layer 2
m_1 terms
n_1 Window Topic documents
A_{DYN}
Topic Modelling
K_{DYN} topics
n_1 documents
W
K_{DYN} topics
m_1 terms
H
Input Matrix (n_1 documents x m_1 terms)
Factor (n_1 documents x k_{DYN} topics)
Factor (k_{DYN} topics x m_1 terms)

Figure 2.8: Dynamic NMF Topic Model - Layer 1 and 2

This two-layer framework works by first applying NMF to individual time periods to produce topics in each time window. In this first layer, the set of speeches being analysed are divided into equally sized disjoint time windows. The disjoint time windows allow for topics to be identified at a point in time in a given window. This caters for the fact that some topics may be short-lived and only span few time windows which would otherwise be hidden if the entire period was analysed using a single time window or overlapping time windows. These individual window topic outputs are then combined by constructing a new matrix, A_{DYN} consisting of the individual rows of the vector H (window topic documents) from all time windows. The top t terms are selected for each window topic and the remaining term weights are set to zero. By selecting the top t terms for each window topic document, it facilitates the inclusion of only the highly descriptive terms in each window topic. The optimal value of t for the second layer of NMF has been found to be 20, as values above this have been found to have little or no impact on the quality of the final dynamic topics generated (Greene & Cross, 2015). A second iteration of NMF is then applied to the new matrix to yield dynamic topics over the entire period being analysed.

Greene and Cross (2015) utilised this approach to identify the agenda of the European parliament at a point in time and over the entire period.

2.8 Validity and Interpretation of Topic Models

2.8.1 Topic Coherence

The semantic validity of the topics is informed by how well each topic can be interpreted as a meaningful cluster. This process can be carried out by reviewing the topics and manually assigning a label to them based on the top terms in each topic and then reviewing a number of randomly selected documents which have a high probability or weighting for that topic (Quinn et al., 2010). Another mechanism for testing validity of the model is by examining the semantic relationship within an individual topic and across groups of topics. Intra-topic validation refers to the validity of a single topic and inter-topic validation measures the relatedness across different topics. The former may be measured using topic coherence which evaluates the degree of semantic similarity across the top words within a topic. This can then be aggregated to get an overall value for all topics within a model. Most of the research in the area of topic coherence focused on probabilistic methods such as LDA. The UCI measure is based on the pointwise mutual information (PMI) (Newman, Lau, Grieser, & Baldwin, 2010) and the score for two words v_i and v_j is calculated as follows:

$$UCIscore(v_i, v_j, \epsilon) = \log \frac{P(v_i, v_j) + \epsilon}{P(v_i)p(v_j)} \tag{2.3}$$

Following this, the UMass measure introduced by Mimno, Wallach, Talley, Leenders, and McCallum (2011) is based on the word co-occurrence frequencies within documents. $D(v_i, v_j)$ is the number of documents which contain words v_i and v_j and the overall score is represented by the following equation:

$$UMassscore(v_i, v_j, \epsilon) = \log \frac{D(v_i, v_j) + \epsilon}{D(v_j)} \tag{2.4}$$

An experiment employed by Stevens, Kegelmeyer, Andrzejewski, and Buttler (2012), which applied the UCI and UMass evaluation measures on an LDA and NMF topic model, indicated that NMF produced less coherent topics than LDA. Further to this study, OCallaghan, Greene, Carthy, and Cunningham (2015) used the opportunity to explore different topic coherence metrics. This research employed three different measures TC-NPMI, TC-LCP and TC-W2V on both NMF and LDA models across a range of corpora. It revealed that NMF is more suited to analysing niche topic areas. LDA was able to produce general descriptions for broad topics but NMF had the capacity to identify more specific topics with high coherence values. Alongside this, Greene and Cross (2015) used topic coherence to measure the validity of the topics produced in the European parliament. It used TC-W2V in conjunction with the word2vec model (Mikolov, Chen, Corrado, & Dean, 2013). TC-W2V is the mean pairwise cosine similarity of two term vectors over a distributed representation for words and is defined as follows:

$$TC - W2V = \frac{1}{\binom{N}{2}} \sum_{j=2}^{N} \sum_{i=1}^{j-1} similarity(wv_i, wv_j) \tag{2.5}$$

This method represents the data in the format of term vectors by using the word2vec model. The word2vec tool consists of two neural network algorithms which are used to estimate word representations in a vector space. It generates word vectors with linguistic regularities from large amounts of text. The extent to which two terms share a similar meaning may be measured by the similarity between term vector pairs. Therefore it may be used in the analysis of topic coherence (OCallaghan et al., 2015). The theory is that topics which consist of highly similar top terms, measured by the similarity between their vector terms, should yield topics which are more semantically coherent.

Therefore, the coherence of an individual topic (t_h) given t top ranked terms to represent it, is given by the mean pairwise cosine similarity between the t corresponding term vectors in the word2vec space as follows:

$$coh(t_h) = \frac{1}{\binom{t}{2}} \sum_{j=2}^{t} \sum_{i=1}^{j-1} cos(wv_i, wv_j) \tag{2.6}$$

The mean coherence of the constituent topics within a model can be used to generate an aggregate model coherence score as conducted by Stevens et al. (2012) and OCallaghan et al. (2015). An aggregate coherence score for the overall model T, comprised of k topics, is calculated by the mean of the individual topic coherence scores as follows:

$$coh(T) = \frac{1}{k} \sum_{h=1}^{k} coh(t_h) \tag{2.7}$$

2.9 Parameter Selection

One area which has not been discussed so far is the input parameters to a topic model which is by and large one of the most important aspects to the implementation of the associated algorithms. The key input value that needs to be decided upfront is the number of topics (k) to be selected. When the value chosen is too small this can lead to very broad topics and if it is too large then there will be too many similar topics being produced. This latter issue is known as 'over-clustering' (Greene, OCallaghan, & Cunningham, 2014). This can be a challenging problem as some data sets have the potential to produce coherent topics at multiple values of k. Greene and Cross (2015) selected an appropriate value of k by running the NMF model over a range of values from k_{min} to k_{max} and calculating the mean model coherence value for each of these applications. The value of k which yielded the highest mean model coherence was selected. Another approach outlined by Greene et al. (2014), using a term-centric stability analysis framework, was based on the principle that a model will be more robust to disruptions in the data when the most suitable number of topics is selected. This work focused on NMF but in theory could be applied to probabilistic topic modeling approaches.

2.10 Strengths and Limitations

One of the main advantages that NMF has over traditional LDA methods is that it has less input parameters that need to be chosen upfront. Also, previous research conducted by Greene and Cross (2017) showed that NMF produces better results than LDA when analysing textual data in specialised domains such as the European parliament. Further to this, OCallaghan et al. (2015) proved that NMF results in more niche topics being identified which may be more generalised in probabilistic approaches. In the political domain, there will be a combination of specialised and broader topics discussed so in theory the algorithm chosen will need to handle a combination of both.

A major limitation to this project is the potential noise within the parliamentary speeches and how best to eliminate this without any negative impacts on the topics being produced. From an initial random review of the content of verbal speeches in the Dáil, there is unnecessarily long speeches and also very short and abrupt content. Also, at this point, it is not clear if the majority of the content of the speeches are specialised or more general. In addition to this, there is a lot of erroneous characters in the data due to the Irish language words being referred to in the English language speeches e.g. Sinn Féin being one of the political parties within the Dáil.

2.11 Summary of Literature Review

This chapter has provided a review of relevant existing literature in relation to the use and importance of PQs across modern countries and also the machine learning techniques and principals required to provide an exploratory platform for the analysis of the Irish PQ data set. An overview of supervisory and unsupervisory methods was given with key examples of algorithms and applications.

Following this, particular emphasis was given to text classification and clustering methods due to the particular data set being analysed. Due to the unavailability of prelabelled data examples and the fact that the ordering of the text within a question is not relevant, topic modeling approaches were reviewed in detail. An overview was given on the theory and applications of LDA and NMF topic modeling algorithms, the former being the most widely used. There exists a wide body of research utilising probabilistic methods and NMF to a much lesser extent and particular attention was given to how they have been applied in niche subject-matter areas. The NMF algorithm produced higher quality topics in niche subject-matter areas over the LDA method (OCallaghan et al., 2015). As such, this study proposes to focus on NMF as the algorithm of choice using the two-layer dynamic model approach introduced by Greene and Cross (2017). Also, as there has been a limited amount of research carried out in the application of NMF in text mining problems, this provided another motivation for using this method. Other areas discussed in this chapter included a dynamic topic model to look at topics over time and approaches used to determine the most appropriate value for the number of topics. As a key area of interest in this study is to see how topics change over time, the two-layer NMF approach will be applied. In addition to this, similar techniques used by Greene and Cross (2017) will be employed for selecting the value for the number of topics and validating the final model results. The reasons for choosing them are twofold. First, these methods have already proven to work on a corpus of political text and second, it was felt that it was more beneficial to spend an extended amount of time on the analyses of the modeling outputs to assess quality of topics generated and highlight trends in political activity.

Non-negative Matrix Factorisation (NMF) in conjunction with the Word2Vec model will determine the model with the most coherent topics for the parliamentary questions in the Dáil. Analysis will be carried out to identify trends between the topics produced and the political parties or politicians that raised them over time.

Finally, to the best of this author's knowledge, there has been no published research in relation to Irish parliamentary questions using unsupervised learning. This research will have the capacity to inform future researchers when applying machine learning methods in the Political Science domain.

The next chapter proposes the design and methodology aimed to tackle the research question.

Chapter 3

Design and Methodology

This chapter details the design steps and methodologies used for implementing the solution and evaluating it. The work being undertaken is a data mining project as it is fundamentally trying to provide an end-to-end solution for analysing large volumes of political textual data and extracting the themes within this data. For this reason, the Cross-Industry Standard Process for data mining (CRISP-DM) (Chapman et al., 2000), as shown in figure 3.1, will be followed where appropriate to plan and design the experiment. However, as this work does not directly involve a business then there are some steps which are not relevant and will be omitted. Remaining sections in this chapter will examine the data sets, provide rationale for making certain decisions with regards to the design and outline any problematic areas encountered in this phase. Finally, this chapter provides a detailed account of each step that will be carried out in the implementation phase.

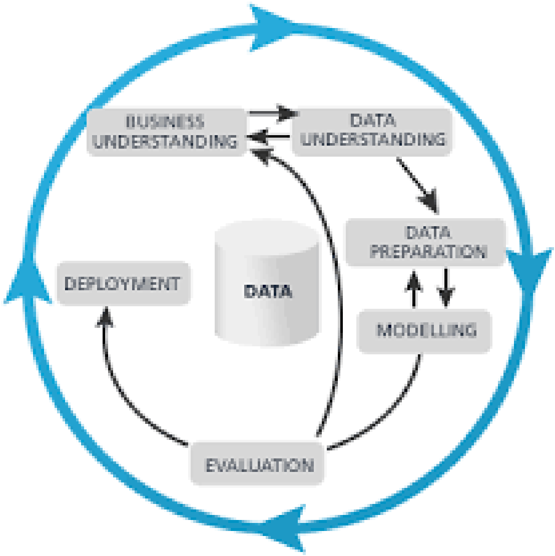


Figure 3.1: CRISP-DM project life cycle(Chapman et al., 2000)

3.1 Overview of Approach/Design

This research crosses over the political science and computer science domains. The core focus is to derive insight in the political science arena using automated techniques and tools within the computer science domain. The project involves the end-to-end processing of large volumes of data using data preparation, predictive modeling and data visualisations techniques and tools to analyse the final results. The specific area of interest is how useful will automated data mining techniques be in providing any insight into the themes of the speeches raised in the Irish Dáil and identifying trends between these themes and the political parties or TDs that raised them over time.

A topic modeling algorithm called NMF will be applied to predict the topics and a coherence measure will be utilised to evaluate and refine the outputs of the model. Finally, the outputs will be manually reviewed, labelled and analysed using data visualisation techniques to identify trends. The speeches data set is unlabelled and hence the need for unsupervised learning techniques. One of the key parameters in topic modeling is the number of topics, k. In order to decide what value to use, the proposal is to compare the coherence of the topics of the model generated for different values of k. The method chosen, proposed by OCallaghan et al. (2015), will use the

TC-W2V measure in conjunction with a word2vec model. The word2vec model will be used to represent the full corpus of input data in a distributional semantic space. TC-W2V will then be used to measure the relatedness of the set of top terms in a topic based on their similarity in this vector space, see equation 2.6. An aggregated value of coherence will be calculated for each value of k tested for each model, see equation 2.7. The model with the highest coherence value will determine the value of k selected. This method will be applied across all time partitions.

To recap, the research question is *"Can a two layer NMF dynamic topic model yield coherent topics in order to highlight trends in the Irish Parliamentary speeches data set?"*

Key steps

The key steps in the design are outlined below. Steps 3-6 can be considered as the basic modeling phase using a predetermined number of window topics and dynamic topics. This facilitates setting up the end-to-end process without being concerned with parameterisations. Following this, steps 7-13 involve more advanced modeling using parameterisations to refine and improve the model outputs.

1. Prepare the data to include data merging, cleansing and partitioning into time windows.

2. Pre-process the prepared documents from each time window to produce TDMs.

3. Apply NMF to each time window to produce the window topic outputs (pre-set number of window topics).

4. Combine outputs from window topics.

5. Apply second layer of NMF to window topic documents to produce dynamic topic outputs (pre-set number of dynamic topics).

6. Convert dynamic topic modeling results into tabular format for review.

7. Build Word2Vec model from entire corpus of input documents across all time windows.
8. Evaluate a range of values for the number of topics k for each time window.
9. Use value of k yielding highest coherence value for each window and apply NMF again to produce new set of window topics for all time windows.
10. Evaluate a range of different values for number of dynamic topics using these window topic documents.
11. Apply second layer of NMF using number of dynamic topics yielding highest coherence value.
12. Review outputs and refine model looking at the stop words and other filtering mechanisms. Re-run models based on parameters chosen.
13. Extract final model results into tabular readable format.
14. Evaluate final topics generated and overall models.
15. Manually review and label the topics from the dynamic model.
16. Load and merge with other datasets.
17. Visualise results and identify trends.

This research is appling a similar methodology utilised by previous research carried out on the European Parliamentary data set which has been published online (Greene & Cross, 2015)[1].

3.2 Data Understanding

3.2.1 Structure of data set

The verbal parliamentary speeches are available to the author in a relational database format. This data set consists of three tables namely speeches, TDs and parties

[1] https://github.com/derekgreene/dynamic-nmf

which are linked by key identifiers. The speeches table contains the speech and other information such as the debate ID, speech date and URL. An additional data set was collected to understand what parties are in government for each sitting Dáil which will be described in the data construction section. The data sets are split into 3 sections for the purpose of how they are analysed in the design phase as follows:

A. Verbal speeches made from 2006 to 2015. [2]

B. Political parties and TDs. [3]

C. Political parties in Government across each sitting Dáil. [4]

The initial step in preparing the data involves analysing the quality of the data and ensuring the integrity of it. Following this, the data will be partitioned over time. Further analysis will be conducted to decide what time partitions are most suitable to the data being analysed. One of the main considerations, which will be part of the pre-processing of the data, will involve the removal of stop-words and filtering of terms within the content of the speeches.

A. Parliamentary Speeches data

The parliamentary speeches dataset is stored in a SQL database. There are 498,161 speeches which span three Dáil terms from 25/01/2006 to 17/12/2015 and 4,225 of these speeches are in Irish with the remaining speeches in English. The Dáil terms are the 29th (06/06/2002 - 30/04/2007), 30th (14/06/2007 - 01/02/2011) and 31st (09/03/2011 - 03/02/2016). Therefore, the speeches being analysed only partially cover the 29th and 31st terms. The speeches relate to 22,427 debates over this period. Each speech relates to a specific debate and there may be multiple speeches made by various TDs within the one debate. The content of these speeches can range from the actual question being raised and subsequent responses or comments made from other parliamentary members. However, not all debates initiate with a specific question but

[2] Sourced from Dr. Derek Greene, University College Dublin

[3] Sourced from Dr. Derek Greene, University College Dublin

[4] http://www.oireachtas.ie/parliament/oireachtasbusiness/parliamentaryquestions/

may commence with a matter raised by the Ceann Comhairle. Each speech is a stream of words which varies in length. In order to process the speeches, they will need to be partitioned into time windows. A sample of this data is shown in table 3.1.

speech_id	debate_id	date_made	langauage	content	URL	td_id
1	2006-01-25_105_0	2006-01-25	en	Question: I would like to ask...	https://www.kildarestreet.com/debate/?id=2006-01-25.106.0	5
2	2006_105_0	2006-01-25	en	This is a critical issue for the irish people.....	https://www.kildarestreet.com/debate/?id=2006-01-25.108.0	7
3	2006-01-25_107_0	2006-01-25_107_0	en	I wish to advise the House of the..	https://www.kildarestreet.com/debate/?id=2006-01-25.113.0	25

Table 3.1: Sample Speeches data set

B Dáil Members dataset

The original dataset retrieved (A and B) has a TD and party name associated with each speech. However, the party which the TD is associated with reflects a recent snapshot in time. There are lots of examples in the last 10 to 15 years where politicians have moved from either Fine Fáil or Fine Gael to independents and vice versa. Some examples of this include Beverly Flynn moving from Fiannna Fáil to Independents and Liam Twomey moving from Independents to Fine Gael. Also, parties which existed some time ago are no longer in existence e.g. Progressive Democrats. For this reason, an additional dataset to replace B is retrieved from the Oireachtas library which depicts all sitting Dáil members and the party they were a member of at the time. This dataset is in tabular CSV format with 4,900 rows which contains all Dáil members (both in government and opposition), their political party, constituency, the Dáil term and house number. The house number represents the house of Oireachtas and is equivalent to the Dáil term as a new house of parliament is elected for each new Dáil term. A sample of this data is shown in table 3.2 below. This new dataset will be merged with the speeches data which will be discussed in the data cleansing section.

Dáil_member_id	td_id	td_name	party_id	party_name	constituency	Dáil_no	house_no
40	3	Bertie Ahern	3	Fianna Fáil	Dublin Central	29th	29
101	3	Bertie Ahern	3	Fianna Fáil	Dublin Central	30th	30
36	48	Aengus Snodaigh	8	Sinn Féin	Dublin SouthCentral	29th	29
115	48	Aengus Snodaigh	8	Sinn Féin	Dublin SouthCentral	30th	30
253	48	Aengus Snodaigh	8	Sinn Féin	Dublin SouthCentral	31st	31
72	73	Beverley Flynn	3	Fianna Fáil	Mayo	29th	29
140	73	Beverley Flynn	6	Independent	Mayo	30th	30
151	205	Aine Brady	3	Fianna Fáil	Kildare North	30th	30
275	284	Aine Collins	2	Fine Gael	Cork NorthWest	31st	31

Table 3.2: Sample Dáil Members data set

C. Political parties in Government

The final piece of data collated is the political parties which were in government and the dates and Dáils that these governments were in place. The purpose of collecting this information is to decipher which politicians were a government or opposition member in Parliament. As per the literature review, this may provide some additional context to the number of speeches being made and also the themes being raised. This information is manually collated from the Oireachtas website and loaded into a table on the database. A sample of this data is shown in table 3.3.

gov_party_id	party_id	party_name	Dáil_no	Dáil_from	Dáil_to
1	3	Fianna Fail	29	2002-06-06	2007-04-26
2	9	Progressive Democrats	29	2002-06-06	2007-04-26
3	3	Fianna Fail	30	2007-06-14	2011-02-01
4	9	Progressive Democrats	30	2007-06-14	2011-02-01
5	7	Green Party	30	2007-06-14	2011-02-01

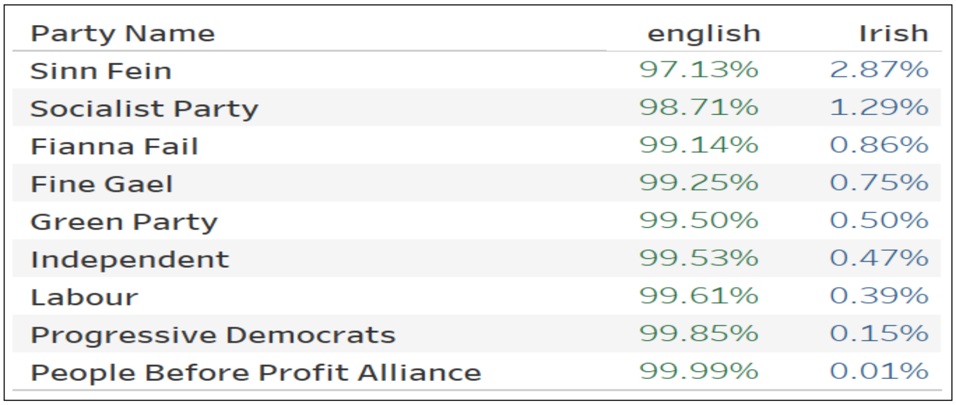
Table 3.3: Sample Parties in Government data set

In summary, the selected data sets are sourced or extracted from the official website/library of the Irish Parliament. They are also deemed substantial in volume and adequate in terms of coverage across Dáil terms for the purpose of the goal of this project. Therefore, it should be possible to build and apply a topic model on this data set for the purpose of extracting the themes within these speeches and identifying trends. As such, a decision is made to utilise the available data set. This in turn means that this study is focusing in on the verbal PQs in the Dáil.

3.3 Data Preparation

3.3.1 Data Cleansing

The speeches data set (A) consists of the verbal parliamentary speeches made from 2006 until 2015 in the English and Irish language. The spread of the Irish language speeches across each party is illustrated in figure 3.2. There is no surprise that the highest percentage of Irish speeches relates to the Sinn Féin party but is less than 3%. As such, by removing all Irish language speeches, this would not detract from the overall study.



Party Name	english	Irish
Sinn Fein	97.13%	2.87%
Socialist Party	98.71%	1.29%
Fianna Fail	99.14%	0.86%
Fine Gael	99.25%	0.75%
Green Party	99.50%	0.50%
Independent	99.53%	0.47%
Labour	99.61%	0.39%
Progressive Democrats	99.85%	0.15%
People Before Profit Alliance	99.99%	0.01%

Figure 3.2: Percentage of speeches made in Irish by political party from 2006 to 2015

For the purpose of this study, a corpus of 493,936 English language speeches is being considered and the 4,225 Irish language speeches made over this time period are excluded. This represents 99.1% of the original speeches dataset. Some relevant statistics on the English speeches data set are illustrated in figure 3.3 which gives the number of speeches across each Dáil term by opposition or government party member.

It can be seen here that the number of speeches is considerably lower in the 29th term and this is due to the fact that the speeches cover only 1.5 years of that 5-year Dáil term. Interestingly enough, this also indicates that the number of speeches made by opposition or government members is not drastically different with government members outweighing opposition party members in the 31st Dáil term.

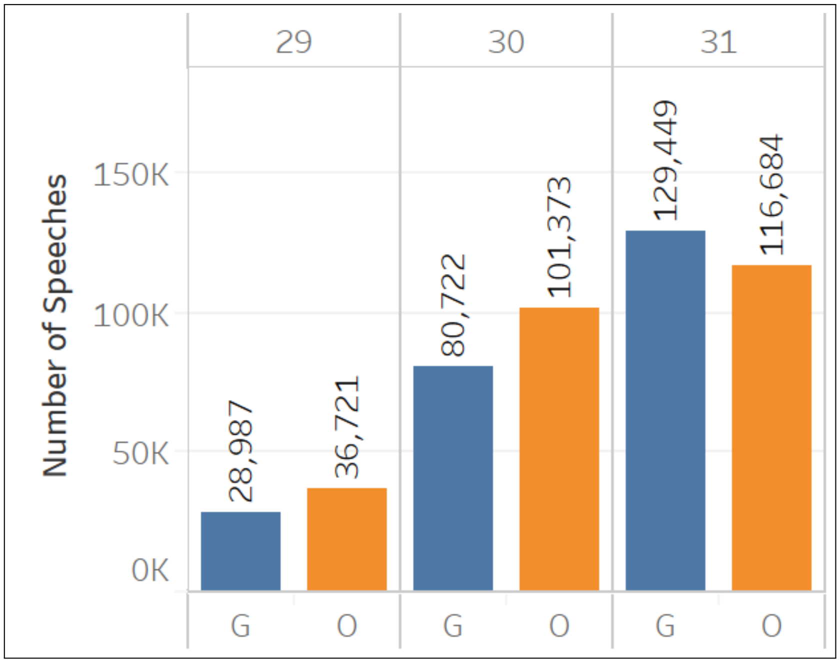


Figure 3.3: Speeches made per Dáil term by opposition or government party members

After an initial review of the speech content, it highlights a number of data quality issues. Firstly, in multiple instances, the content of the speech is truncated due to a parsing error caused by the presence of special characters. Further to this, it is noticed that there are erroneous characters present where there are accents, or as known in Irish as a ”fada”. Examples of such words found are ”Sinn Féin”, ”Fianna Fáil”. Other issues encountered on the speech itself are apostrophes and dashes. These issues are handled by deleting the erroneous characters or replacing them with the correct characters where possible.

Once the quality of speech content is confirmed, the next step is to review the integrity of the links between the speeches and the Dáil members. As previously mentioned, the original data set provided did not account for the movement of politicians between parties and represented a snapshot in time. The new data set retrieved (B) needs to be merged with the existing speeches data. Adequate analysis and testing is completed to ensure that the speeches linked accurately to the new Dáil members data set. During this merging process, a number of data quality issues are identified. Firstly, the names of the TDs on the new data set versus the existing speeches did not always align, so special care had to be taken to ensure that they were accurately linked. Alongside this, there are issues with politicians in the same family (father and son) having the same name but sitting on the Dáil at different times. Both of these issues are dealt with manually by reviewing the records and correcting the links. It was also identified that a couple of TDs had no existing td_id which meant that they sat on the Dáil but had never given a speech over the period. Finally, there are TDs which are not listed as Dáil members which appear to have given speeches on the 90th anniversary of the Dáil. For the purpose of this study, the speeches involving the latter two issues are excluded. In summary, a lot of effort is spent ensuring the accuracy and integrity of the data as part of this data preparation phase. This process was a worthwhile exercise as it means that each speech can now be tracked against each Dáil member and their associated party at any point in time.

3.3.2 Data Construction/Partitioning

A new dataset is constructed for the parties in Government. This lists all parties in government across the 29th, 30th and 31st Dáils from 2006 to 2015. The Dáil term is added to the speeches dataset based on when the speech is made. The number of words in the content of each speech is calculated to allow for analysis of the input data. Finally, two additional variables are derived which includes a field to identify whether a TD is in government or opposition and whether a speech is the lead speech in the debate. Both of these fields will be used in the final analysis of the results.

The next step in the data preparation phase is deciding how best to split the data into time windows over the ten-year period in order to facilitate the dynamic aspect to the topic model. This is a key design decision. As the analysis is following a 2-layer NMF approach the 493,936 speeches will initially be partitioned into time windows to produce window topics and then a second layer of NMF will produce dynamic topics over the entire period. The distribution of speeches is analysed over a number of time intervals namely year, quarter and month. The volume of speeches by year is illustrated in figure 3.4.

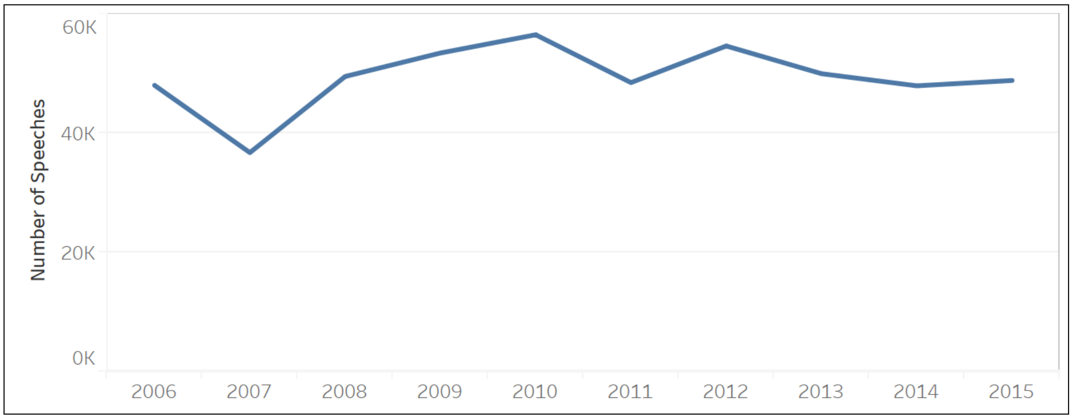


Figure 3.4: Speeches made per year

At a quarterly level, as illustrated in figure 3.5, the highest number of speeches is generally made in Q4 of each year and the lowest in Q3. This can be explained by the fact that the Dáil summer recess takes place for 12 weeks and as a result of this break, there is a flourish of activity in Q4 such as budget time. Looking further at a monthly level, as per figure 3.6, it is clear that it would not be suitable as there are very little speeches made in August and September due to the recess period. After consideration of the spread of the data across various time intervals and previous work carried out by Greene and Cross (2015), it is decided to partition the data into equal distinct quarterly time windows over the ten-year period. This yields 40 distinct non-overlapping time windows from Q1 2006 to Q4 2015. Based on this partitioning, the lowest number of speeches in any one of these time windows is 2,486 in Q3 2007 (consisting of a corpus of 257,834 words) to a maximum of 20,058 speeches in Q4 2008. Overall, across the full period, the average number of words per quarter is 1.4 million and the average number of words per speech is 119.7. These facts provide evidence that there are adequate number of speeches in all 40 time windows to perform the 2-layer topic modeling approach.

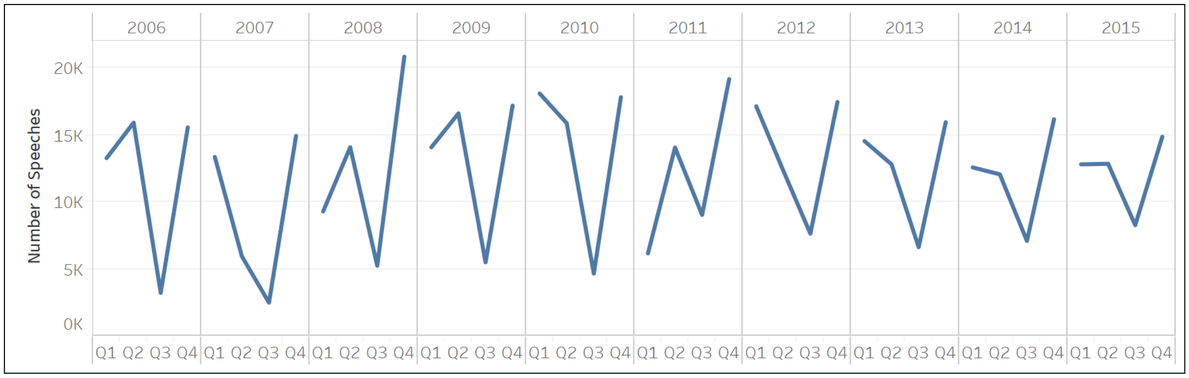


Figure 3.5: Speeches made per quarter per year

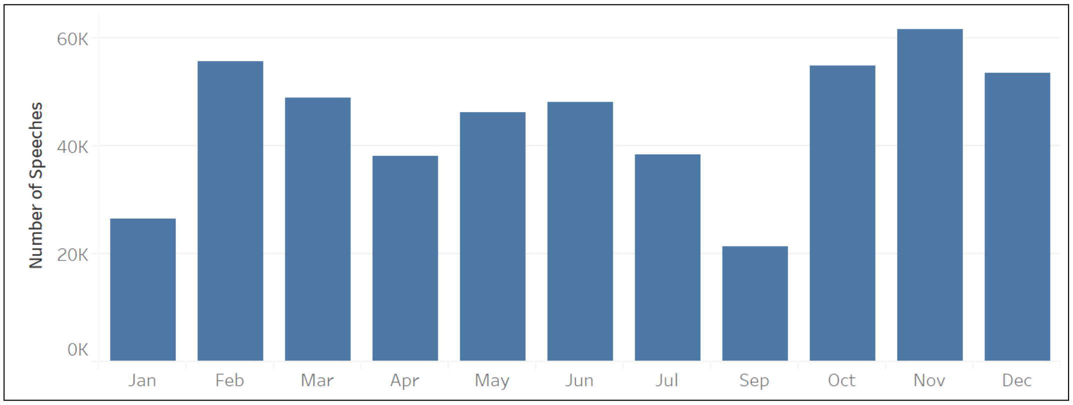


Figure 3.6: Speeches made in monthly category across full period 2006 -2105

In order to process the speeches, they will be exported from the speeches table and converted into individual .txt files.

3.3.3 Data Pre-processing

At this point, it is necessary to consider the terms within the speech text and what level of pre-processing should be undertaken to improve the modeling results. This will involve text mining processes such as lemmitization, tokenisation and filtering.

The first step in the modeling process is to pre-process the inputs so that they are in a suitable format to perform the topic modeling. For each quarterly time window, a document-term matrix, A_t, is constructed. This involves the following steps for each time window:

1. Select all speeches from that time window.
2. Tokenize each document.
3. Remove short tokens with less than 3 characters.
4. Remove common stop-words.
5. Remove parliamentary related stop-words e.g. 'taoiseach'. This list is compiled after the first few model iterations by reviewing what parliamentary terms are coming up in the top 50 terms for each cluster. Also, after each iteration, new terms will be added to this list as necessary. The criteria for adding a term to the list will involve judgment by the author as to whether a specific term provides context to a particular cluster/topic.
6. Remove tokens occurring in less than 0.1% of all speeches. A very small value was initially chosen similar to approach used by Greene and Cross (2015).
7. Build matrix $\mathrm{A_t}$ based on remaining tokens.
8. Execute TF-IDF term weighting and document length normalisation. This weighting factor has been illustrated to work well when the goal is to produce both broad and niche topics (OCallaghan et al., 2015).

These steps will be repeated 40 times to produce 40 TDMs for each time window $\mathrm{T_i}$.

3.4 Modeling

This section will describe in detail the steps involved in building the two-layer dynamic topic model. These steps will be run multiple times to improve the quality of the topics being generated. One of the main reasons for this is to ensure that the appropriate number of topics is selected for the both the window topic models and dynamic topic model over a range of values tested. Initially, an arbitrary value for k is selected for both the window and dynamic topics in order to set up the end-to-end model.

This value is set at 7 for all such models. Following this, the experiment to test the coherence value across a range of values of k is conducted.

3.4.1 Modeling Layer 1 NMF

Once the speeches across all the time windows have been pre-processed into their TDMs, the first layer of NMF will be applied to these matrices to produce the window topics. The output to this stage will be 2 vectors, a document-topic vector and a topic-term vector for each window stored in a .pkl file. Note, at this point, an arbitrary value is chosen for the number of topics for each time window.

3.4.2 Modeling Layer 2 NMF

After creating the window topics, the resulting window topic documents will be combined and used as the inputs to the 2nd layer of the NMF model to produce the dynamic topics. Again, in this case, the number of topics will be set at an arbitrary value. The output to this stage will be one file containing the dynamic topics to window topic documents factor and dynamic topics to terms factor.

3.4.3 Build Word2Vec model

The Word2Vec model will be built from the entire corpus of speeches from 2006 to 2015.

3.4.4 Evaluation of Number of Topics for each Time Window

For this step, topic coherence will be the measure to evaluate the number of topics for each window. This provides a measurement to evaluate the association of the set of top ranked terms in an individual topic. This is calculated by comparing the similarity of the top terms in the cluster against the distributed semantic space compiled by the word2vec model. TC-W2V, as defined in equation 2.6, is the coherence measure calculated here. Then using equation 2.7, an aggregated mean value of this coherence measure can be calculated for each model within each time window. This experiment will be conducted by first running the first layer of NMF for each time window across a range of values for the number of window topics, k_{min} to k_{max}. The range tested will be from 4 to 40 and the output will be the top 3 values for k with the highest level of coherence at the overall model level. This recommendation can differ across time windows and will lead to different values of k being selected across time windows.

3.4.5 Rerun layer 1 NMF

Based on the recommended number of topics for each time window, the first layer of NMF is run again to produce a new set of window topic documents.

3.4.6 Evaluation of Number of Dynamic Topics

The window topic documents generated from the previous step are combined and a 2nd layer of the NMF is run to test a range of values for the number of dynamic topics from 4 to 40. This test is based on the same principle as the window topics test using the TC-W2V coherence measure calculated for the top terms represented in the word2vec semantic space. The highest coherence measure calculated at the model level for each value tested will result in the number of dynamic topics being selected.

3.4.7 Review Outputs and Refine Model using Parameterisations

The terms in each dynamic topic will be reviewed and interpreted. At this stage, the parameters set as part of the data pre-processing stage are revisited i.e. general stop-words, list of parliamentary stop-words, thresholds for the percentage of all speeches that the terms must appear in. Finally, other filtering options are explored depending on the quality of the outputs. It was clear from the initial executions of the dynamic topic model that the top terms being generated for both dynamic and window topics provided little or mixed context to their meaning. Some topics contain a lot of very general terms. In order to get any meaning from them, a number of refinements as listed below are made across a given time window and specified value for the number of topics.

The time window chosen for this refinement stage is Q4 2008. This window is chosen as it has the largest number of speeches. Before any processing is complete, it contains 20,958 speeches which represent almost 5% of all English language speeches. Furthermore, in order to ensure that other conditions are kept stable, a value of k (number of topics) is selected. This was achieved by using the optimal value of k for that window which has been selected to this point (k=20) and it is kept static during this pre-processing evaluation phase.

1. The list of stop-words is expanded to include not only common stop-words such as 'the','of' etc. but also to include common political terms used in the Dáil such as 'taoiseach', 'tanaiste', 'statement' and the names of all politicians who have been a member of the Dáil. In total, this list contained 348 common stop-words and common political terms. The set of all TD's first and surnames were added to this list. The final list contained 837 stop-words in total.

2. For a given time window, terms are removed which occurred in less than or greater than a specified number of speeches. This input parameter was altered from a small range to large range to see how the outputs would be impacted. For example, for a term to be included it must appear in 10% to 60% of all documents within a time window. However, this range was deemed too narrow so these attributes were refined for a particular window until the outputs were improved.

3. The other and perhaps most important aspect of the pre-processing phase of the data was experimenting with more aggressive filtering options. It became apparent after executing the end to end process a number of times that the topics being generated were very difficult to interpret. Therefore, it was decided to explore the filtering of the data by extracting the nouns only. This included the extraction of general nouns and proper nouns only from the content of the speeches. The latter specifies people, places, things, or ideas. This decision had a significant positive

impact on the interpretability of the clusters generated, therefore, was applied to the speech content.

4. The final additional step taken in this pre-processing phase was the replacement of plural words with their singular version. The need for this step became apparent after the nouns had been extracted. Note, it was important that the plural term was replaced and not removed so that it still contributed to the overall frequency count of the term it was representing.

The results, as per table 3.4 include only the executions after the nouns are filtered as they are deemed the most relevant.

	Time Window	**k (window topics)**	**Stopwords**	**Min df**	**Max df**	**Additional Filtering**	**Coherence value**	**Selected**
1	Q4 2008	20	837	0.1	80	Nouns and Pronouns only left in with plural replacement	0.3376	N
2	Q4 2008	20	837	0.1	90	Nouns and Pronouns only left in with plural replacement	0.3538	Y
3	Q4 2008	20	837	0.1	99	Nouns and Pronouns only left in with plural replacement	0.3536	N

Table 3.4: Parameter Selection Results for Pre-processing step

To give a sense of the impact of the filtering parameters, the original 20,858 speeches for Q4 2008 are reduced down to 16,441 documents after option 2 in table 3.4 is applied. Of the 4,417 documents that are filtered out, 4,197 of them have speeches which contain less than 10 words. A sample review is conducted of the 220 speeches which contained greater than 10 words to confirm the validity of this exclusion. Once the pre-processing step is reapplied to each time window, it results in 20% of speech documents being removed. The remaining 392,892 speech documents are used to compile the input TDM for each time window and the modeling process is run again including the step to select the values of k (number of topics).

3.4.8 Reformat and Load Modeling outputs

The final outputs will produce a distribution of ranked weights for each term and for each speech against each window topic. Similarly, there will be a distribution of ranked weights for each term and each window topic against each dynamic topic generated. For the purpose of this study, a single membership model will be employed which means that a document can relate to only one topic but multiple documents can relate to the same topic. This is based on the highest ranked weight that a document has against the topics generated. This principle is used to extract the final window model results and dynamic model results into a readable format. This output will contain the relationship and ranked weightings between the dynamic topics to window topics to speeches and their underlying top terms.

3.4.9 Identify Labels for Dynamic Topics

Based on the top terms, the final topics generated will be manually reviewed and labelled appropriately. The top ten terms should provide the context to a topic, if it is not clear by the top ten terms then further terms may be reviewed.

3.4.10 Visualise Final Results

Finally, the topics generated are analysed across time, party, TD and Dáil term to identify potential trends. The number of speeches that a politician has made, which is associated with the window topics that make up the dynamic topic, will provide knowledge as to whether they had an interest in a particular topic. The same analysis will be completed by party and will be represented over time. Other fields will also be reviewed such as constituency and the opposition/government flag.

3.5 Evaluation Methods

3.5.1 Window Topic Modeling Evaluation

The window topics and overall window topic model evaluation is two-fold. First, the coherence measure calculated using the appropriate value of k for each time window may be assessed by looking at the average value of coherence across all topics for a given time window model. This will provide validation if the model has produced coherent topics. Secondly, each window topic selected will have top terms associated with it. A review of a selection of these topics with their top terms will provide meaning as to what it relates to. Also, each window topic will have the top speech documents which relate to it so a review of these will provide a means for confirmation of the topics generated.

3.5.2 Dynamic Topic Modeling Evaluation

A similar evaluation will be made for the dynamic topic model by looking at the average coherence across all topics in the model. If this value proves to be above the lower bound limit for the coherence measure then it can be accepted that the results are of good quality. Further to this, a manual review will be carried out on the top terms in each dynamic topic and also the top terms in the window topics which each dynamic topic relates to for further verification.

3.6 Strength & Limitations of the Solution/Approach

The main strength behind this design is the fact that it is utilising a methodology which has proven to be successful on another data set in the political science domain. To counteract this somewhat, the European parliamentary data set used in previous research may have presented more curtailed discussions than those speeches allowed in the Irish Parliament so this does display itself as a potential concern. For example, there are often matters discussed in the Dáil which are trivial such as the dress code. Very often there are outbursts from TDs such as the recent one between the Haely Rae family and Marc MacSharry over the new question time system. Such dramatics would not have presented itself in the European Parliament.

As preliminary research uncovered, verbal questions are sometimes used to reach wide audiences due to the amount of media coverage they get above written questions. This may mean that a lot of very general topics could be generated. To cater for this concern, the TDMs for each time window are initialised with a TD-IDF term weighting factor. The process of applying this weighting factor to the data prior to the application of the NMF model has been illustrated to assist the production of varied but semantically coherent topics which are less likely to be represented by same top terms (Greene & Cross, 2015). This allows NMF to identify both general and specialised topics.

Repeating this experiment with different pre-processing conditions is an important key aspect to this design which significantly improves the quality of the topics generated. Also, by running the NMF model over a very wide range of values for the number of topics ensures that the optimal value of k is always selected. A single metric (TC-W2V) was calculated as the coherence measure. To further substantiate the findings, additional coherence measures could have been utilised.

Another aspect to this design worth mentioning is the single membership model. This allows for a document to be related to only one topic based on its highest weighting across all topics. This will mean that some documents may have a very high weighting for a topic and others may have a much lower one and there is no differentiation made on these variances in this study.

Further to this, the data was partitioned into 40 equal quarterly time windows. There are other ways that this could have been performed such as by splitting the data by each sitting government to produce window topics for each one. However, quarterly partitions would have a higher number of partitions thus providing the greater potential to uncover topics which were short-lived and discussed in few windows.

Chapter 4

Implementation, Evaluation and Results

This chapter details the results and the evaluation of them. It will commence by giving a summary of the implementation details which were not discussed as part of the design chapter. It will outline in detail the experimental results and the final analysis of the modeling outputs with other relevant data from the political science domain. It will discuss in detail any trends comparing them to external factors at the time and also highlighting any limitations of them. A summary of these findings will conclude this chapter.

4.1 Introduction

The previous chapter gave a detailed account of the design and methodology. To summarise the design steps, they may be grouped into 5 main tasks.

1. Pre-processing

2. Modeling and refinement

3. Evaluation.

4. Manual review of topics and identify appropriate labels.

5. Visualisation of final results and trends.

The first two tasks have been described in detail in chapter 3 so this chapter will focus on the final three stages.

4.2 Key Implementation details

4.2.1 Data Pre-processing

The implementation of the data pre-processing steps was carried out using C# and python scripts. The former was used to extract the speeches and compile the text documents within time window folders. Each speech is timestamped as to when the speech is made so each text file will be labelled according to the time window it relates to and its unique identifier on the database i.e. speech ID. This is to ensure that there is a link from the model outputs back to the original dataset for analysis of the results. Python was used for the tokenisation, stop-word filtering and generation of the TDMs in the format of .pkl files.

4.2.2 Modeling

As per the literature review, NMF was the chosen topic modeling algorithm. The models were built in Python using, where possible, scripts available online[1]. The set of scripts use the numpy, scikit-learn, prettytable and genism packages made available via PIP. Some of these scripts were refined to make them more appropriate to the problem and data set at hand.

This modeling process generates a number of outputs. For each time window model, it generates a python pickle file for each value of k that is tested (4 to 40). Each pickle file generated consists of a window topic to term matrix and window topic to speech matrix for all 40 windows. These matrices are extracted from the pickle file and converted into readable CSV formats. A similar set of files is generated for the 2nd layer of the dynamic model. Again, python scripts are used to extract these

[1] https://github.com/derekgreene/dynamic-nmf

factors and the final single membership model results into CSV formats. A C# script is used to load the final model results into the original database for analysis purpose. Finally, these results are analysed and visualised using Tableau.

4.3 Results

4.3.1 Modeling Results

The first layer of the dynamic model produced a set of 494 static window topics. The top 20 terms for these window topic documents are then used to generate the input matrix to the second layer of the model. This generates a matrix of 494 window topics and 6,026 distinct terms. The second NMF layer applied produced 26 dynamic topics. This can be interpreted as 26 different topics discussed over the entire 10-year period. Looking at the number of static topics generated across each time window, as per figure 4.1, the number of topics selected against number of speeches within that time window have a similar pattern with some exceptions. One quarter, which has a small number of speeches (4,165) versus a high number of topics (30), is Q3 2008. Comparing this to the external factors at the time, the small number of speeches coincides with the holiday recess period and the high number of topics could be accounted for by the emerging economic crisis in Ireland. This would have impacted different sectors, thus explaining the high number of topics discussed at that point in time.

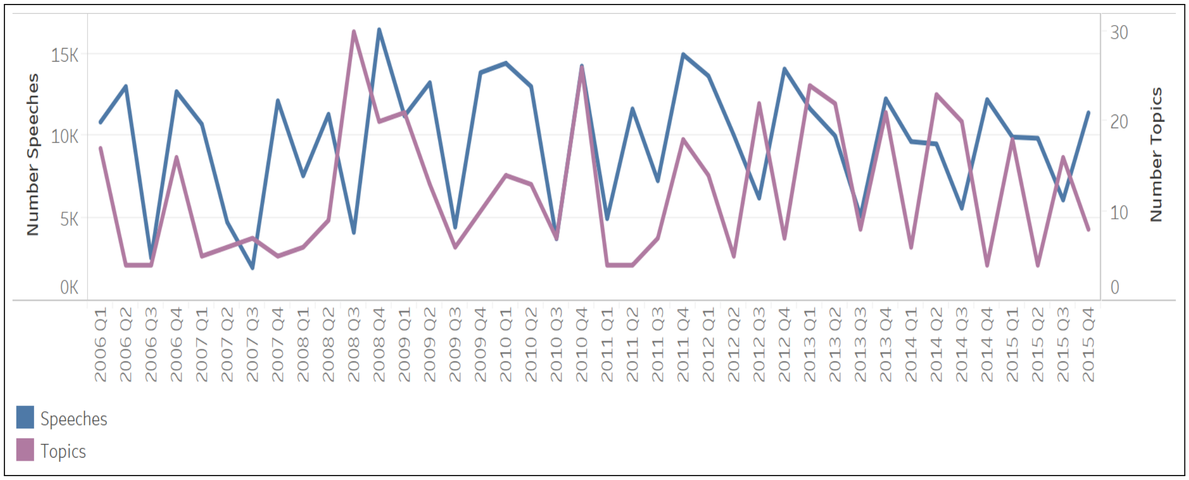


Figure 4.1: Number of Topics versus Number of Speeches for each time window

4.3.2 Evaluation of Results

The modeling results are evaluated in two manners. First, the coherence of the topics and overall model are calculated. This measures the relatedness between terms within a topic. It gives a mathematically measure for confirmation of the quality of the topics generated for both the window topic models and the dynamic model. To recap from the literature review, this process is known as intra-topic validation which tests the meaningfulness of the top terms within an individual topic. The coherence measure used is TC-W2V and is calculated at the topic and overall model level using equations 2.6 and 2.7 on page 27. Following that, specific examples of the dynamic topics generated are analysed in detail to provide further confirmation of them. This involved reviewing the underlying speeches that relate to a particular dynamic topic and also comparing the trends found with external factors at the time.

Window Topic Modelling

To recap, the first layer of the NMF model is applied 36 times for each static time window which equates to 1,440 iterations across all 40 time windows. Using the coherence measure, the most appropriate value for the number of topics (k) is selected for each window as illustrated in figure 4.2. These values range from 4 to 30.

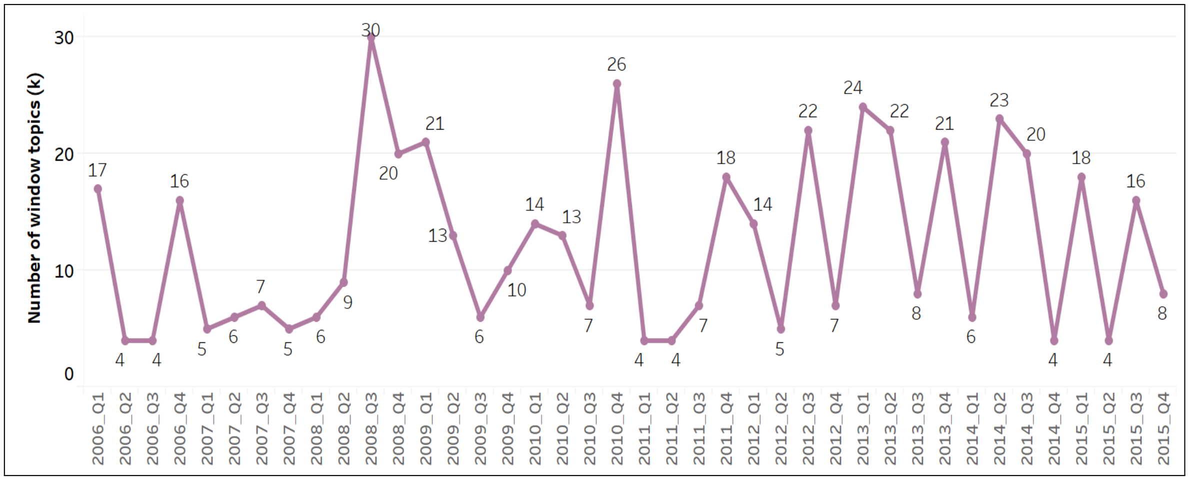


Figure 4.2: Number of topics selected for each time window

The coherence values against each time window, based on the value of k selected from figure 4.2, are illustrated in figure 4.3. This is the average topic coherence for each model produced for each time window based on the coherence of the underlying topics. This shows that the coherence of the window topic model ranges from 0.3247 to 0.3764 across all 40-time windows. Overall, the mean topic coherence minimum value of 0.3247 is considerably above the lower bound for TC-W2V (-1.0). This suggests a high level of semantic validity across all window topic models.

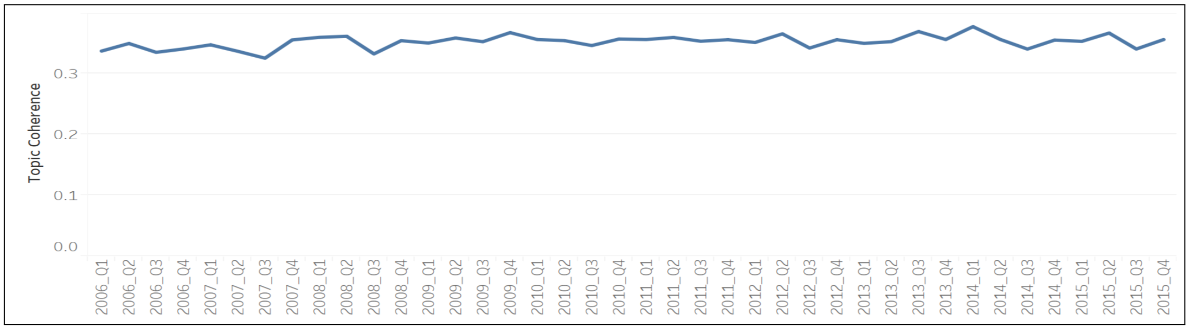


Figure 4.3: Topic Coherence across each time window

Dynamic Topic Modelling

The dynamic topic model is applied 36 times to select the optimal value of k, from 4 to 40. The optimal value of k for the dynamic topic model is found to be 26. This provides an intra-topic validation measure which is the mean topic coherence score across all dynamic topics of 0.3736. This value is significantly higher than the lower bound limit of TC-W2V (-1.0) and less than the upper limit (1.0) which illustrates semantic validity across the dynamic topic model.

4.3.3 Manual Review & Labelling of Topics

The coherence measure for the overall dynamic topic model and the minimum coherence value for all window topic models demonstrates a high level of semantic validity across these models. In order to further investigate the validity of the topics a manual review is carried out and labels are identified. There are 26 dynamic topics generated over all time. This means that these topics were discussed in one or multiple time windows. The manual review involved assigning a label to each topic based on their top ten terms. These labels and top terms of the niche topics only are illustrated in table 4.1. This process shows that 16 clusters relate to specific topics and the remaining 10 are very broad and relate to order of the day/policy/general matters. Specific topics generated include health, education, employment, water charges, child and family, financial crisis, Ireland & EU, environment/heritage, social care, FOI requests, justice, legislation, housing, taxes and finance, reports commissioned.

Label	Term 1	Term 2	Term 3	Term 4	Term 5	Term 6	Term 7	Term 8	Term 9	Term 10
Budget/finance	budget	finance	cut	plan	welfare	figure	measure	expenditure	decision	spending
Child/Family	child	family	parent	care	right	woman	support	need	youth	home
Education	school	education	teacher	student	pupil	need	science	project	level	building
Employment	job	company	economy	enterprise	sector	employment	creation	investment	worker	plan
Financial crisis	bank	mortgage	central	irish	anglo	credit	debt	banking	guarantee	money
FOI requests	information	freedom	request	office	body	record	public	commission	commissioner	respect
Guards/Crime	garda	justice	equality	commissioner	siochana	law	reform	crime	force	station
Health	health	service	hospital	care	patient	insurance	bed	emergency	system	staff
Heritage/Environment	government	environment	decision	policy	commitment	heritage	election	reform	whip	motion
Housing	housing	authority	dublin	council	rent	list	project	county	city	transport
Ireland & EU	ireland	country	decision	union	eu	position	european	agreement	policy	respect
Legislation	legislation	law	piece	provision	head	content	court	item	regulation	respect
Report commissioned	report	implementation	group	commission	progress	publication	justice	review	woman	inquiry
Social care	person	home	country	family	system	life	welfare	work	money	society
Taxes	tax	property	income	rate	finance	charge	revenue	labour	relief	household
Water	water	charge	service	household	supply	meter	authority	conservation	cost	infrastructure

Table 4.1: Top ten terms in the niche dynamic topics generated (16 topics)

At this point, the terms within the overarching dynamic topics and window topics have been reviewed manually. There are 494 window topics which make up the 26 dynamic topics, therefore the number is too high to run through in detail. However, in order to illustrate the quality of the underlying window topics, the top ten terms within the window topics for a sample of dynamic topics are represented in figures 4.4, 4.5, 4.6. These are word clouds illustrating the top ten words in the underlying window topics that make up the dynamic topics. The size of the word is based on the frequency of occurrence of the term in the top ten terms in the underlying window topic documents that constitute the dynamic topic. This shows that the top terms within the underlying window topics are related to the dynamic topic's subject matter.

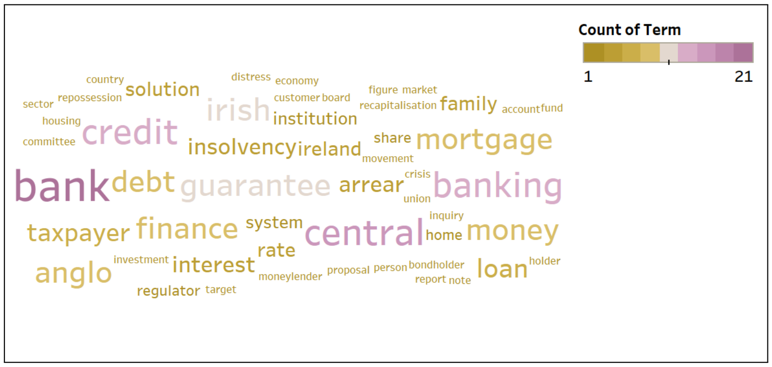


Figure 4.4: Financial Crisis by top terms within related Window Topics

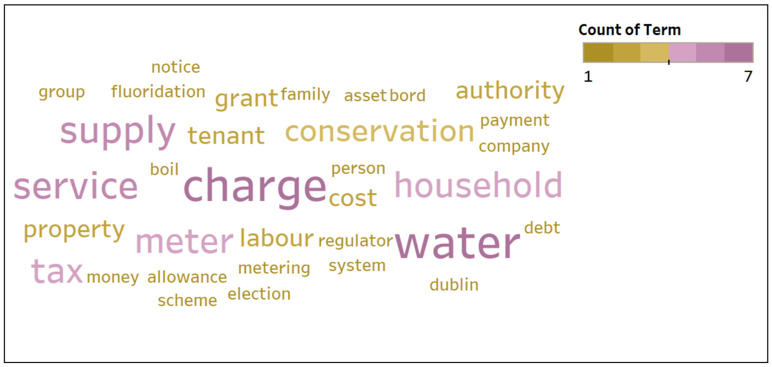


Figure 4.5: Water by top terms within related Window Topics

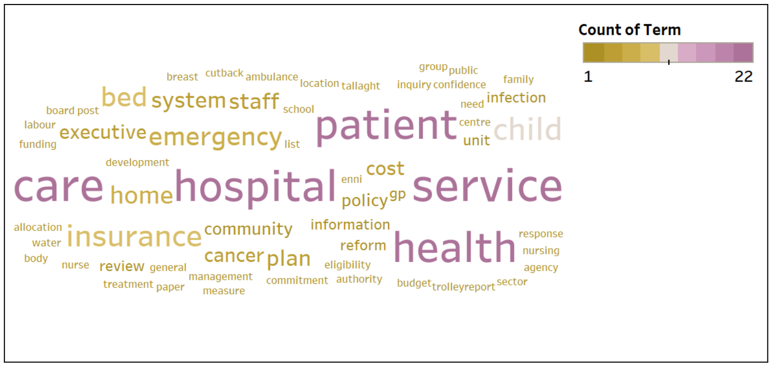


Figure 4.6: Health by top terms within related Window Topics

As previously stated in the design chapter, a single membership model is chosen which in turn means that a speech relates to only one window topic and each window topic relates to only one dynamic topic. This relationship is based on the highest ranked weighting of a document against a topic which is output in factor W for both layers of the NMF model. Therefore, to illustrate the speeches in an underlying topic, table 4.2 shows the top ranked speech across 6 of the 21 time windows that the financial crisis was discussed in. The content of these speeches provides further validation of this cluster. The full list of top ranked speeches in all 26 time windows is available in the appendix.

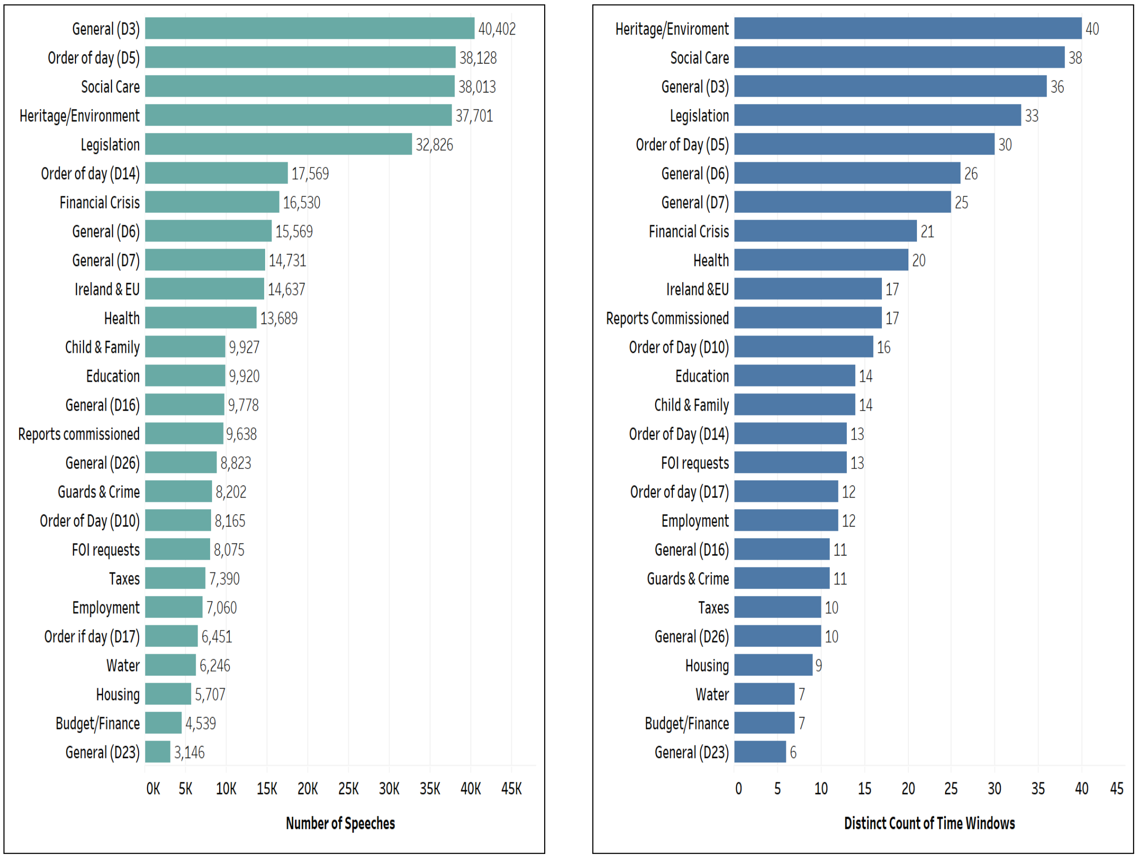
Speech ID	Dail	TD	Speech Excerpt
123104	30	John O'Donoghue	Asking on the Order of Business who pays if a bank goes under -come on Deputy Gilmore.
139602	30	Brian Cowen	The Anglo Irish Bank which is the bank to which the Deputy referred.
170628	30	Michael Ahern	In the 1970s I worked as a financial controller in a large building firm in Cork. At that time the country was in the throes of an economic crisis not much different and possibly worse than that being experienced today. I know first hand the problems..
206085	30	Michael D. Higgins	Anglo Irish Bank the systemic bank.
208512	30	Sean Fleming	I welcome the opportunity to speak on the Central Bank Reform Bill 2010. I want to deal with the specifics of the legislation and then outline the further measures that will be needed to regulate Irish banks beyond the contents of the legislation. After..
292651	31	Enda Kenny	We called in the banks and had a discussion with them about lending policy in general and whether AIB and Bank of Ireland would be in a position to meet the 2011 target of lending 3 billion apiece 3.5 billion next year and 4 billion in 2013. We...

Table 4.2: Financial Crisis - sample of top ranked speeches in 6 quarters

4.4 Trend Analysis

4.4.1 Topic priorities overall in the Dáil

Following the labelling of the 26 topics, they were analysed in detail. The first area that was looked at was the importance of these topics overall by understanding how frequently they were discussed and what was there longevity over the analysis period. Figure 4.7a depicts the number of speeches made for each topic and figure 4.7b illustrates their occurrence across the quarterly time windows. As expected, general/order of day matters are discussed the most with the highest volume of speeches. To clearly focus in on the trends across the niche topics, all general topics are grouped together under one heading. Therefore, all graphs from this point forward will illustrate 18 topics instead of 26 i.e. 16 niche topics and one general and order of the day topics.

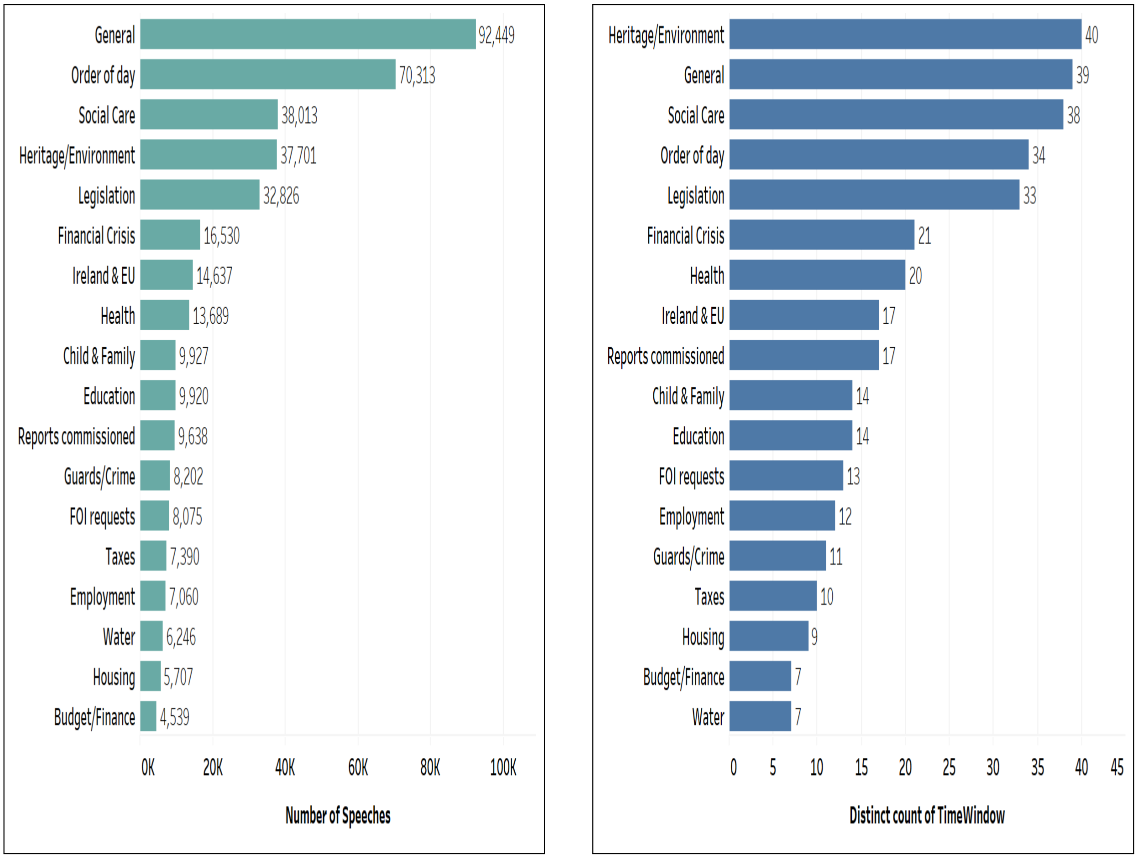


(a) Topics by number of speeches

(b) Topics by distinct quarters

Figure 4.7: Dynamic topics importance over ten year period

Figure 4.8a and figure 4.8b illusrate the number of speeches and the number of quarters the topic was discussed using this new topic grouping. From this, the prime niche topics discussed are social care, heritage/environment and legislation which aligns to the niche topics which have the highest longevity occurring in 33 to 40 quarters. Issues relating to the financial crisis, Ireland & the EU and health have a high number of speeches and are discussed in more than 17 quarters. Topics such as the budget, housing and water issues have the lowest number of speeches respectively which again aligns to the fact that they are discussed in fewer quarters over the analysis period.



(a) Topics by distinct quarters (Grouped)

(b) Topics by distinct quarters (Grouped)

Figure 4.8: Dynamic topics importance over ten year period

It is also worth ackowledging that a large number of speeches have been clustered into very general topics and order of the day items. This is an area which could be investigated further to see if it was possible to extract more meaningful context/themes from them.

4.4.2 Topic priorities by party and constituency

Looking at the number of speeches per topic and per party, figure 4.9 demonstrates what political parties are discussing. The parties and topics are both sorted by total number of speeches in descending order. Excluding general/order of the day matters, the top ranked niche topics by volume of speeches are social care, heritage/environment and legislative matters for all parties. Note, the Progressive Democrats have been excluded from this graph as they are no longer in existence. There are similar trends across the two biggest parties, Fianna Fáil and Fine Gael. Contrary to the bigger parties, Independents and the Sinn Féin party have shown to prioritise matters relating to water, taxes, guards/crime and housing. Also, the top three topics being discussed by the socialist party are heritage/environment, social care and water issues. Finally, the Green party has remarkably no speeches in relation to water issues and the socialist party having the next lowest number of speeches for the financial crisis topic.

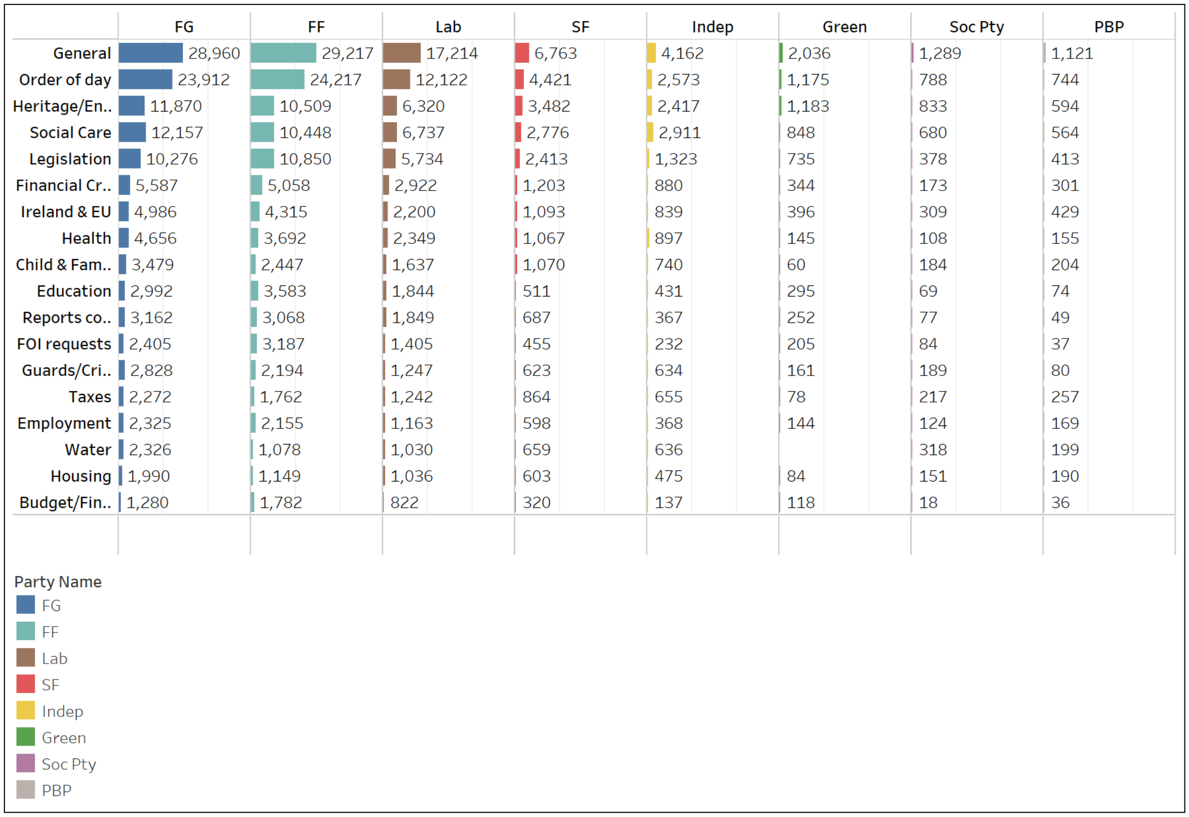
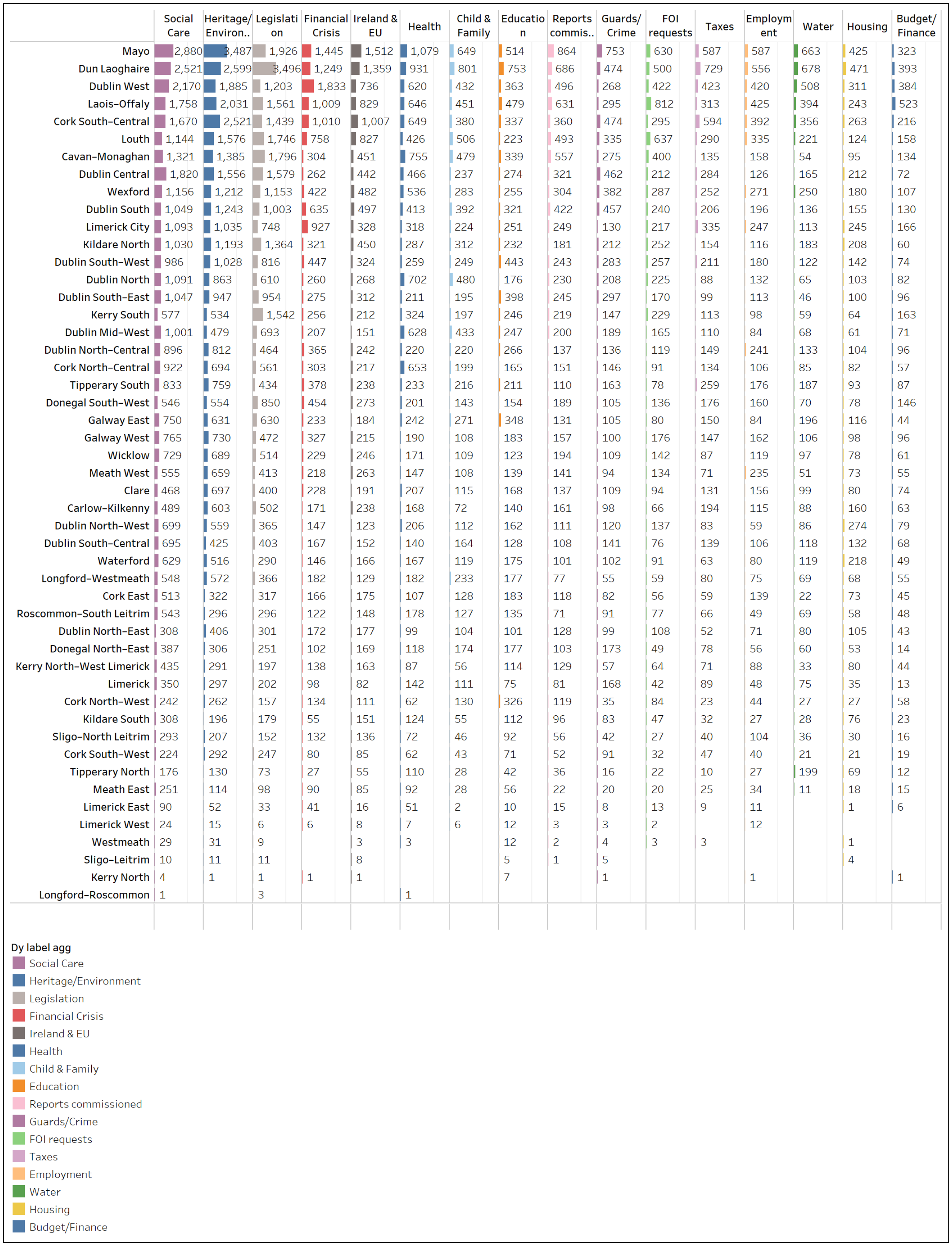


Figure 4.9: Dynamic topics by number of speeches and by party

Figure 4.10 illustrates the number of speeches per topic per constituency that a TD represents. This graph depicts the 16 niche topics only in order to focus in on these areas only. Both categories have been ordered by volume of speeches. One would expect the overall number of speeches per constituency to be related to the number of representative seats in those areas. For example, Mayo where Enda Kenny is from and Dublin-West would have a high number of government ministers and spokespersons for the opposition. However, it does highlight what topics have been given priority by TDs in those constituencies. For the constituencies at the higher end of the scale they are discussing legislation, social care, heritage/environment and financial crisis issues in the main. This trend is similar to what was seen at both a party and overall level. It does illustrate a few exceptions to this such as Tipperary North discussing water issues the most; Meath-East and Cork North-Central having an interest in health matters. Also, there have been 6 constituencies at the lower end of the scale which have not discussed all of the topics. This could be down to TDs in these areas being interested in more local or general matters or perhaps not having any knowledge on these subject matters. In particular, Limerick East, Limerick West, WestMeath, Sligo-Leitrim, Kerry North and Longford-Roscommon have not had any speeches relating to water issues.

4.4.3 Further detailed analysis on specific topics

A sample of the topics generated in table 4.1 are selected for further review and validation. The specific topics are selected based on 3 factors, namely, how much they have been discussed, their distribution across time windows and finally, matters of public interest as identified by the author. The rationale is to select dynamic topics which have been discussed by all parties which spread across a large number of time windows versus a small number. With this in mind, the financial crisis, water, health, Garda/crime and housing are chosen for further analysis.



	Social Care	Heritage/ Environ..	Legislation	Financial Crisis	Ireland & EU	Health	Child & Family	Education	Reports commis..	Guards/ Crime	FOI requests	Taxes	Employment	Water	Housing	Budget/ Finance
Mayo	2,880	3,487	1,926	1,445	1,512	1,079	649	514	864	753	630	587	587	663	425	323
Dun Laoghaire	2,521	2,599	3,496	1,249	1,359	931	801	753	686	474	500	729	556	678	471	393
Dublin West	2,170	1,885	1,203	1,833	736	620	432	363	496	268	422	423	420	508	311	384
Laois–Offaly	1,758	2,031	1,561	1,009	829	646	451	479	631	295	812	313	425	394	243	523
Cork South–Central	1,670	2,521	1,439	1,010	1,007	649	380	337	360	474	295	594	392	356	263	216
Louth	1,144	1,576	1,746	758	827	426	506	223	493	335	637	290	335	221	124	158
Cavan–Monaghan	1,321	1,385	1,796	304	451	755	479	339	557	275	400	135	158	54	95	134
Dublin Central	1,820	1,556	1,579	262	442	466	237	274	321	462	212	284	126	165	212	72
Wexford	1,156	1,212	1,153	422	482	536	283	255	304	382	287	252	271	250	180	107
Dublin South	1,049	1,243	1,003	635	497	413	392	321	422	457	240	206	196	136	155	130
Limerick City	1,093	1,035	748	927	328	318	224	251	249	130	217	335	247	113	245	166
Kildare North	1,030	1,193	1,364	321	450	287	312	232	181	212	252	154	116	183	208	60
Dublin South–West	986	1,028	816	447	324	259	249	443	243	283	257	211	180	122	142	74
Dublin North	1,091	863	610	260	268	702	480	176	230	208	225	88	132	65	103	82
Dublin South–East	1,047	947	954	275	312	211	195	398	245	297	170	99	113	46	100	96
Kerry South	577	534	1,542	256	212	324	197	246	219	147	229	113	98	59	64	163
Dublin Mid–West	1,001	479	693	207	151	628	433	247	200	189	165	110	84	68	61	71
Dublin North–Central	896	812	464	365	242	220	220	266	137	136	119	149	241	133	104	96
Cork North–Central	922	694	561	303	217	653	199	165	151	146	91	134	106	85	82	57
Tipperary South	833	759	434	378	238	233	216	211	110	163	78	259	176	187	93	87
Donegal South–West	546	554	850	454	273	201	143	154	189	105	136	176	160	70	78	146
Galway East	750	631	630	233	184	242	271	348	131	105	80	150	84	196	116	44
Galway West	765	730	472	327	215	190	108	183	157	100	176	147	162	106	98	96
Wicklow	729	689	514	229	246	171	109	123	194	109	142	87	119	97	78	61
Meath West	555	659	413	218	263	147	108	139	141	94	134	71	235	51	73	55
Clare	468	697	400	228	191	207	115	168	137	109	94	131	156	99	80	74
Carlow–Kilkenny	489	603	502	171	238	168	72	140	161	98	66	194	115	88	160	63
Dublin North–West	699	559	365	147	123	206	112	162	111	120	137	83	59	86	274	79
Dublin South–Central	695	425	403	167	152	140	164	128	108	141	76	139	106	118	132	68
Waterford	629	516	290	146	166	167	119	175	101	102	91	63	80	119	218	49
Longford–Westmeath	548	572	366	182	129	182	233	177	77	55	59	80	75	69	68	55
Cork East	513	322	317	166	175	107	128	183	118	82	56	59	139	22	73	45
Roscommon–South Leitrim	543	296	296	122	148	178	127	135	71	91	77	66	49	69	58	48
Dublin North–East	308	406	301	172	177	99	104	101	128	99	108	52	71	80	105	43
Donegal North–East	387	306	251	102	169	118	174	177	103	173	49	78	56	60	53	14
Kerry North–West Limerick	435	291	197	138	163	87	56	114	129	57	64	71	88	33	80	44
Limerick	350	297	202	98	82	142	111	75	81	168	42	89	48	75	35	13
Cork North–West	242	262	157	134	111	62	130	326	119	35	84	23	44	27	27	58
Kildare South	308	196	179	55	151	124	55	112	96	83	47	32	27	28	76	23
Sligo–North Leitrim	293	207	152	132	136	72	46	92	56	42	27	40	104	36	30	16
Cork South–West	224	292	247	80	85	62	43	71	52	91	32	47	40	21	21	19
Tipperary North	176	130	73	27	55	110	28	42	36	16	22	10	27	199	69	12
Meath East	251	114	98	90	85	92	28	56	22	20	20	25	34	11	18	15
Limerick East	90	52	33	41	16	51	2	10	15	8	13	9	11		1	6
Limerick West	24	15	6	6	8	7	6	12	3	3	2		12			
Westmeath	29	31	9		3	3		12	2	4	3	3			1	
Sligo–Leitrim	10	11	11		8			5	1	5					4	
Kerry North	4	1	1	1	1			7		1			1			1
Longford–Roscommon	1		3			1										

Figure 4.10: Niche dynamic topics by number of speeches and by constituency

Topics evolvement over time

Figure 4.11 graphs how the 5 selected topics evolve over time based on the number of speeches made. From this graph, it is evident that health and the financial crisis topics have both been discussed over a long period and housing and water less so. Also, it can be seen that the highest number of speeches made across all 5 topics in any one quarter relate to water despite it being a short-lived topic.

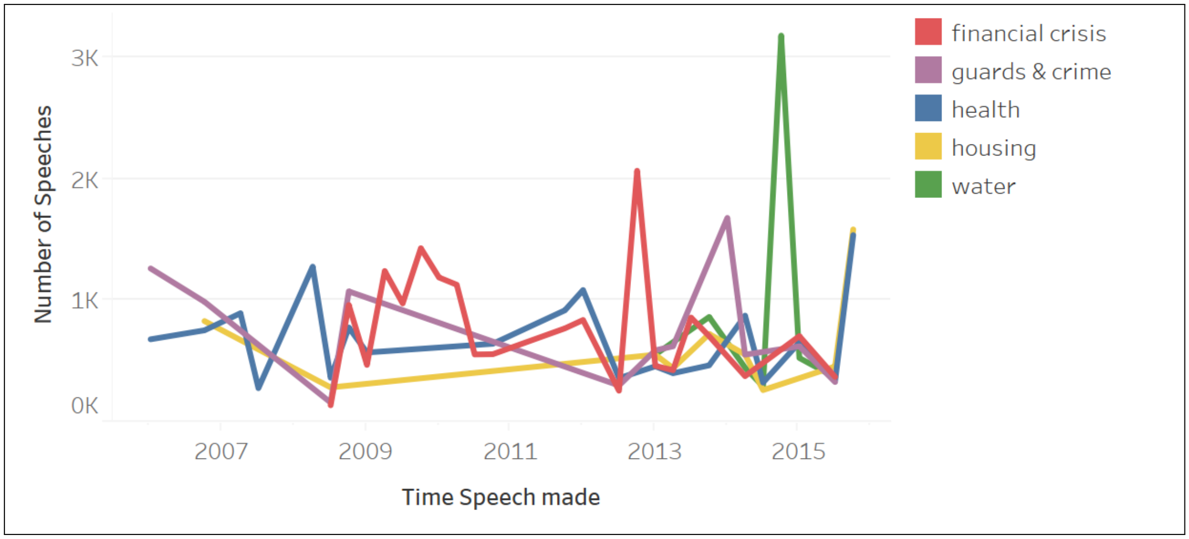


Figure 4.11: Dynamic topics by number of speeches (5 selected)

The financial crisis topic is spread across 21 of the 40 time windows with a total of 16,532 speeches. Figure 4.12 illustrates how this topic has evolved over time by the number of speeches and whether the speech was made by a government or opposition member of the Dáil. Interestingly, the first window that it is raised in the Dáil is in Q3 2008. This is when the Minister for Finance agreed to issue a state guarantee on the Irish banks which was the defining moment of the economic crisis in Ireland. The fact that the timeline for this dynamic topic mirrors the factual basis to when the financial crisis would have been debated is a validation of this clustering process. There is no real difference in the number of speeches made by government or opposition members with the exception of Q4 2009 whereby opposition has made 60% of the speeches for this topic in this time window. This suggests opposition party members interrogating the government over some of the decisions made and the impact on the Irish people.

The dynamic topic relating to water is distributed across only 7 of the 40 time windows. This fact is illustrated in figure 4.13 which shows how the number of speeches made by opposition or government members have evolved over time. There is a significant higher number of speeches in 2014 Q4 which ties in with a large protest against water charges which happened on 11th of October that year when over 50,000 people marched in Dublin. Overall, there are 6,246 speeches over the period. This is a prime example of a topic which could have been hidden if the two-layer dynamic topic modelling approach had not been adopted.

Figure 4.14 illustrates health which has been discussed in 20 out of the 40 time windows. This is no surprise given the ongoing issues within the Irish Healthcare system that this topic would be discussed over a long period. The highest number of speeches (1,541) pertaining to health are in Q4 2015.

Figure 4.15 illustrates the Guards/crime topic which has been discussed in 11 out of the 40 time windows. Over the past decade, there were on-going issues with An Garda Siochana. However, some of these issues regarding specific commissions of investigations, such as the Higgins report, have been clustered into a separate topic under the topic 'Reports Commissioned'. The highest number of speeches made (1,683), in relation to the Guards/crime, is in Q1 2014. This aligns to when the Garda Commissioner at the time, Martin Callinan, resigned in March 2014 as a result of information released by whistleblowers, Maurice McCabe and John Wilson.

Figure 4.16 illustrates housing which has been discussed in 9 out of the 40 time windows. The highest number of speeches made (1,588) is in Q4 2015. This is a low number of quarters that housing has been discussed in, given the impact of the financial crisis on the housing sector since 2008. However, matters relating to the housing sector such as mortgages arrears have fallen in under banking matters under the financial crisis topic which can be illustrated in the samples speeches for this topic in the appendix.

In summary, on review of the number of speeches made by opposition versus government members, there is no major difference in these and are illustrated to be fairly well balanced when looking at individual topics.

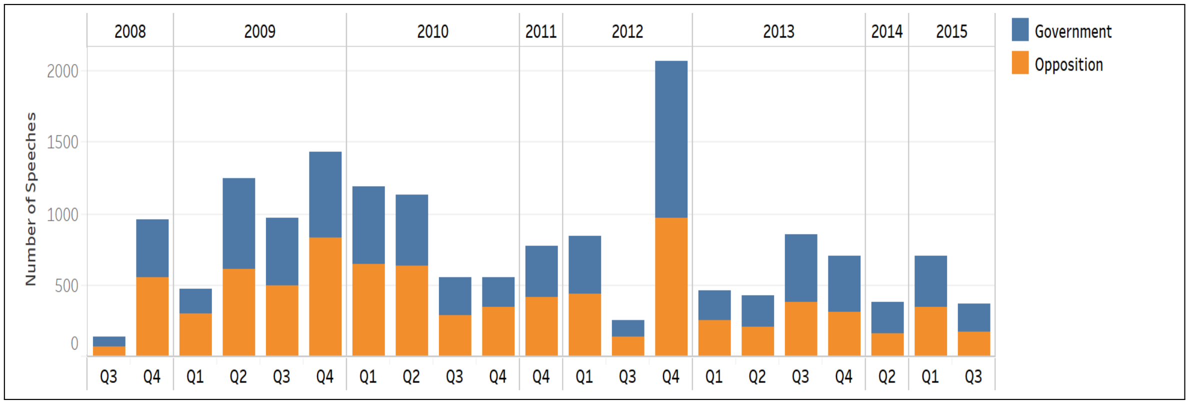


Figure 4.12: Financial Crisis - Number of speeches by government/opposition member over time

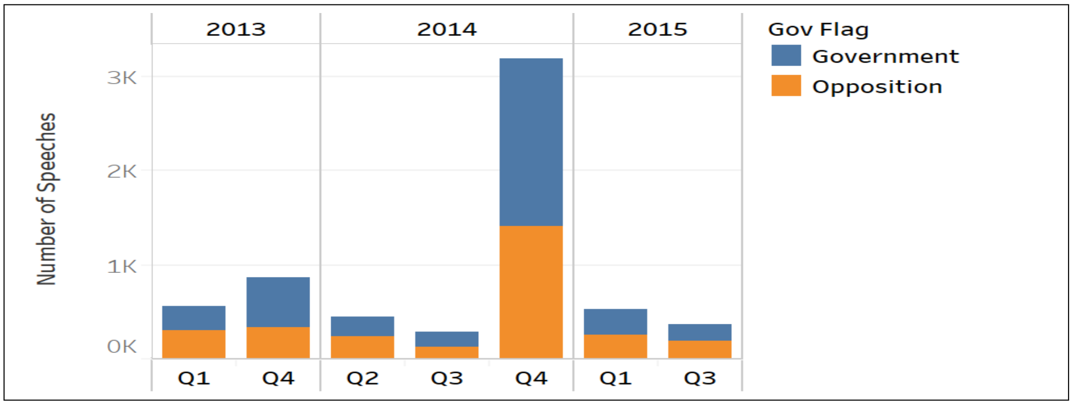


Figure 4.13: Water - Number of speeches by Year

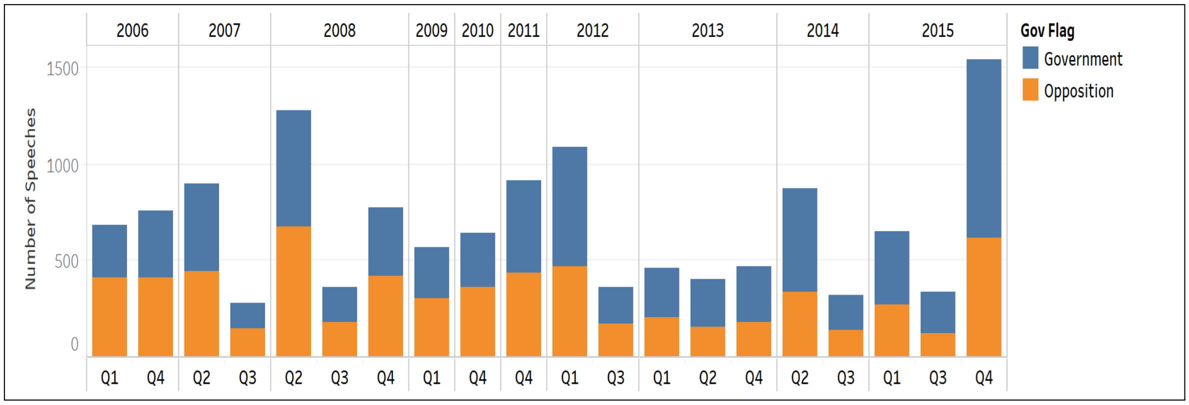


Figure 4.14: Health - Number of speeches by Year

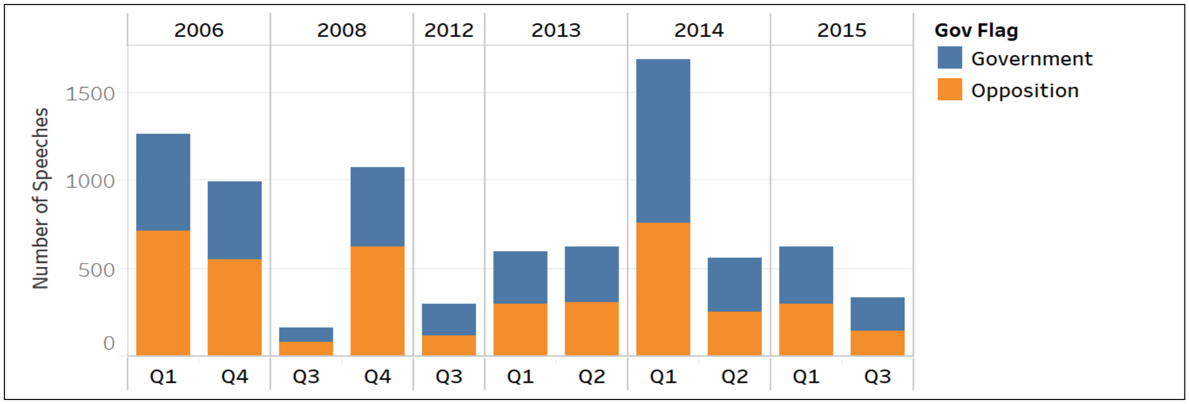


Figure 4.15: Guards/Crime - Number of speeches by Year

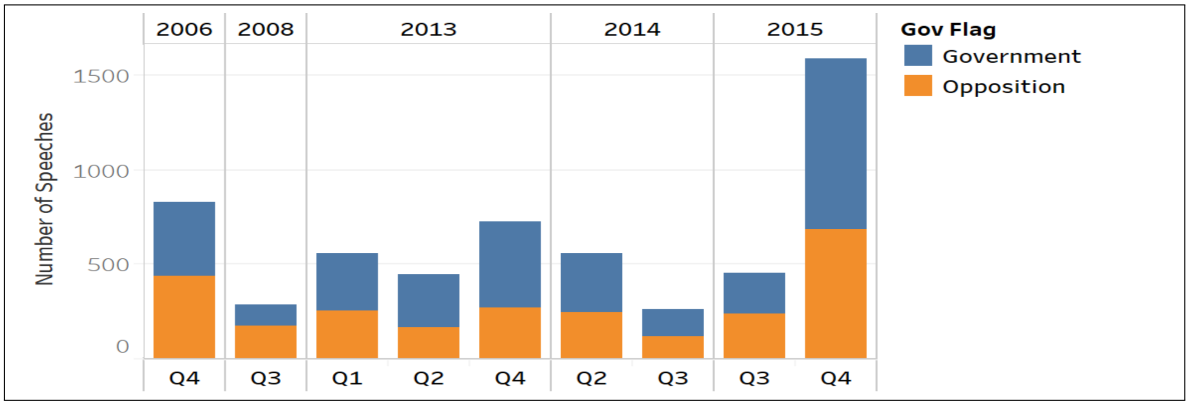
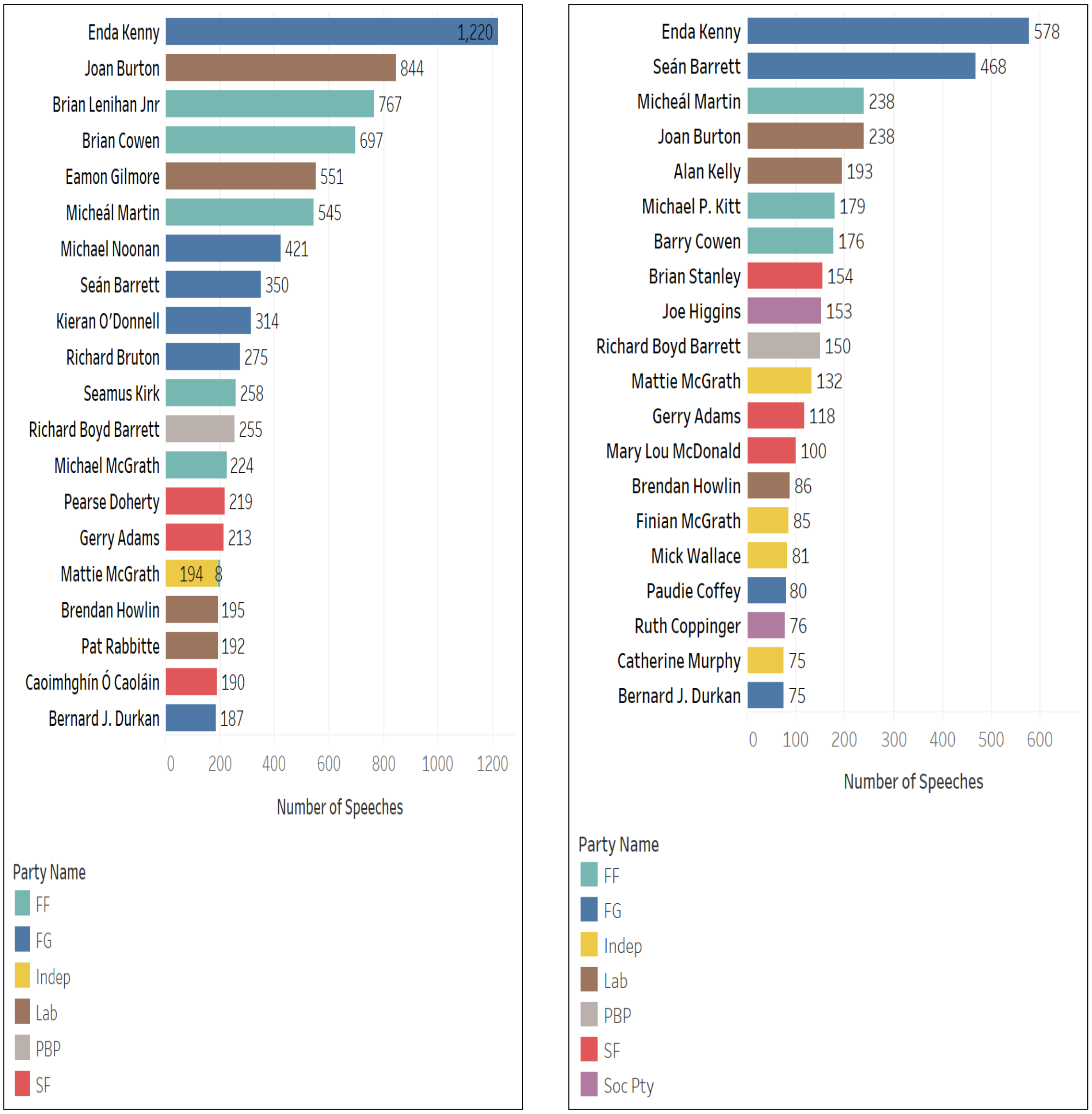


Figure 4.16: Housing - Number of speeches by Year

Topics by TD

Figure 4.17a illustrates the top 20 TDs which have been discussing the financial crisis by volume of speeches. It can be seen that the main party leaders are in the top 20 TDs involved in these debates with Enda Kenny, as head of the Fine Gael party, having the highest number of speeches followed by Joan Burton, the leader of the Labour party. Brian Cowen, the Taoiseach in place when the crisis hit, and Brian Lenihan, as Minister for Finance, both representing the Fianna Fáil party, are also in the top 5. It is also worth noting that Sean Barrett and Seamus Kirk served as Ceann Comhairle (chairperson of the Dáil) over the period in question and are therefore in the top 20 list. Note, Mattie McGrath is appearing as a stacked bar in figure 4.17a which represents his membership in the Fianna Fáil party until 2011 and as an Independent

TD since then.



(a) Financial Crisis

(b) Water

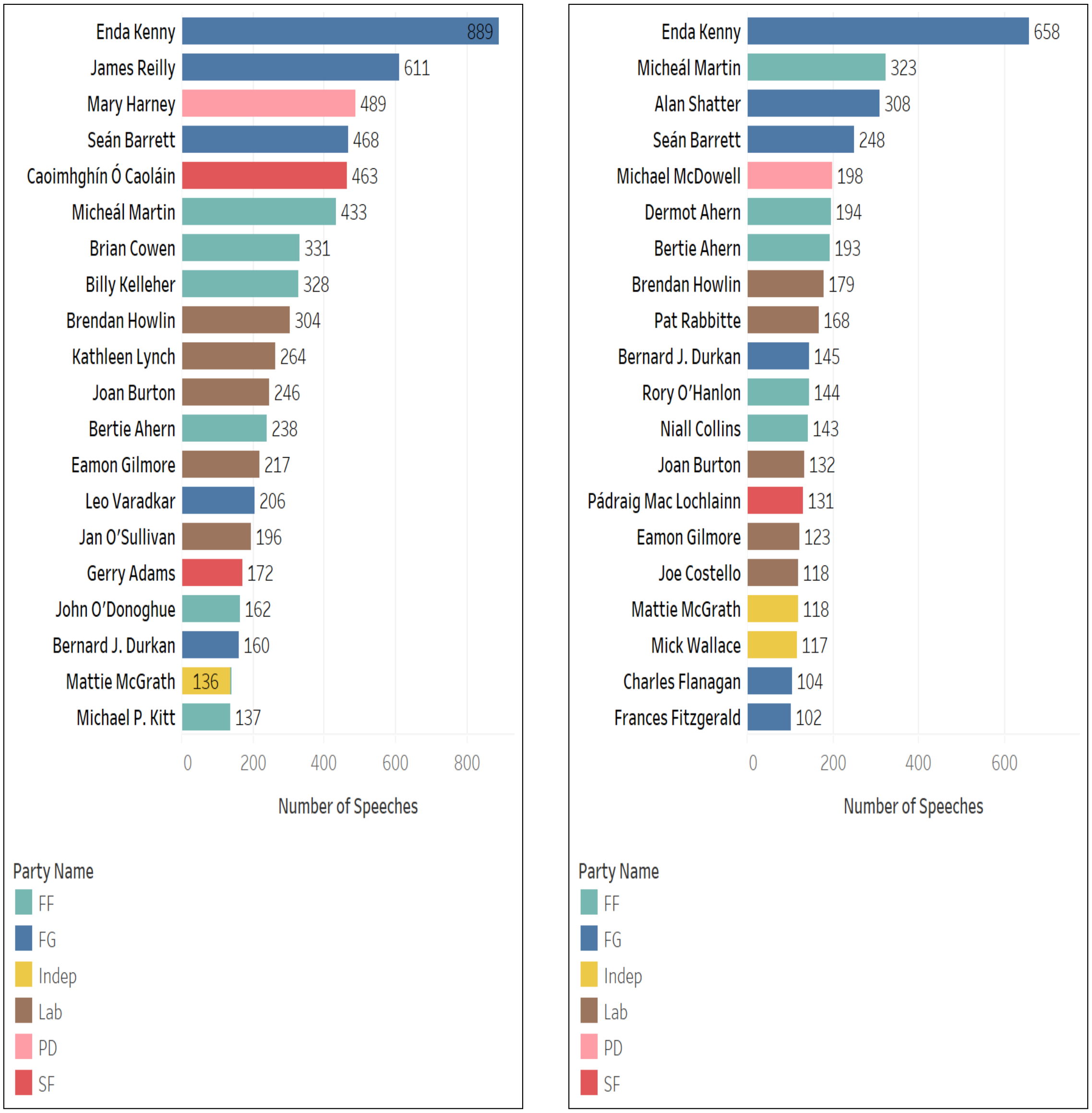
Figure 4.17: Topic: TD by number of speeches (Top 20)

Figure 4.17b depicts the top 20 TDs, by number of speeches, which have discussed water issues. This list includes the main party leaders and the Ceann Comhairle. Alongside these, Alan Kelly, whom was the Minister for environment that introduced the water charges and Barry Cowen, the Fianna Fáil environmental spokesperson, appear in the top seven. The independents are well represented with four TDs appearing in this list. Richard Boyd Barret, a member of the People Before Profit party, would also have been very active in this area.

The top TDs discussing health matters, by number of speeches, are depicted in figure 4.18a. Enda Kenny, Mary Harney and James Reilly are the top 3 speakers for health issues. This is understandable given that Mary Harney served as leader of the Progressive Democrats party and Minister for Health from 2004 to 2011. James Reilly succeeded Mary Harney as Minister for Health and remained in that position until July 2014. There is only one independent TD discussing health in the overall top 20 which is similar to the financial crisis topic and in both of these cases the TD is Mattie McGrath. Contrary to this, there are 5 independent TDs in the top 20 speeches for water issues.

Figure 4.18b illustrates the top 20 TDs discussing Guards/crime matters by the number of speeches. The main party leaders along with the Ministers for Justice, Alan Shatter, Michael McDowell, Dermot Ahern and Frances Fitzgerald all appear in this list. Charles Flanagan as Fine Gael spokesperson on Justice, Equality and Reform from 2007 to 2010 has made a large contribution to this topic. The independent TDs showing an interest in this area are Mattie McGrath and Mick Wallace. It is unusual that Gerry Adams, leader of the Sinn Féin party, is not appearing in this list and there is only one TD from this party present.

The contribution of TDs to housing matters is depicted in figure 4.19. Jan O'Sullivan, as Minister of State for Housing and Planning from 2011 to 2014, appears in the top five along with Fine Gael party leader and the Ceann Comhairle. Richard Boyd Barrett is appearing very high on this list, which may be due to his campaigns against high rise developments and the bank-bail out, which had a major impact on the housing sector. Again, Mick Wallace is showing an interest in this topic followed by Catherine Murphy from the independents.



(a) Health

(b) Guards/Crime

Figure 4.18: Topic: TD by number of speeches (Top 20)

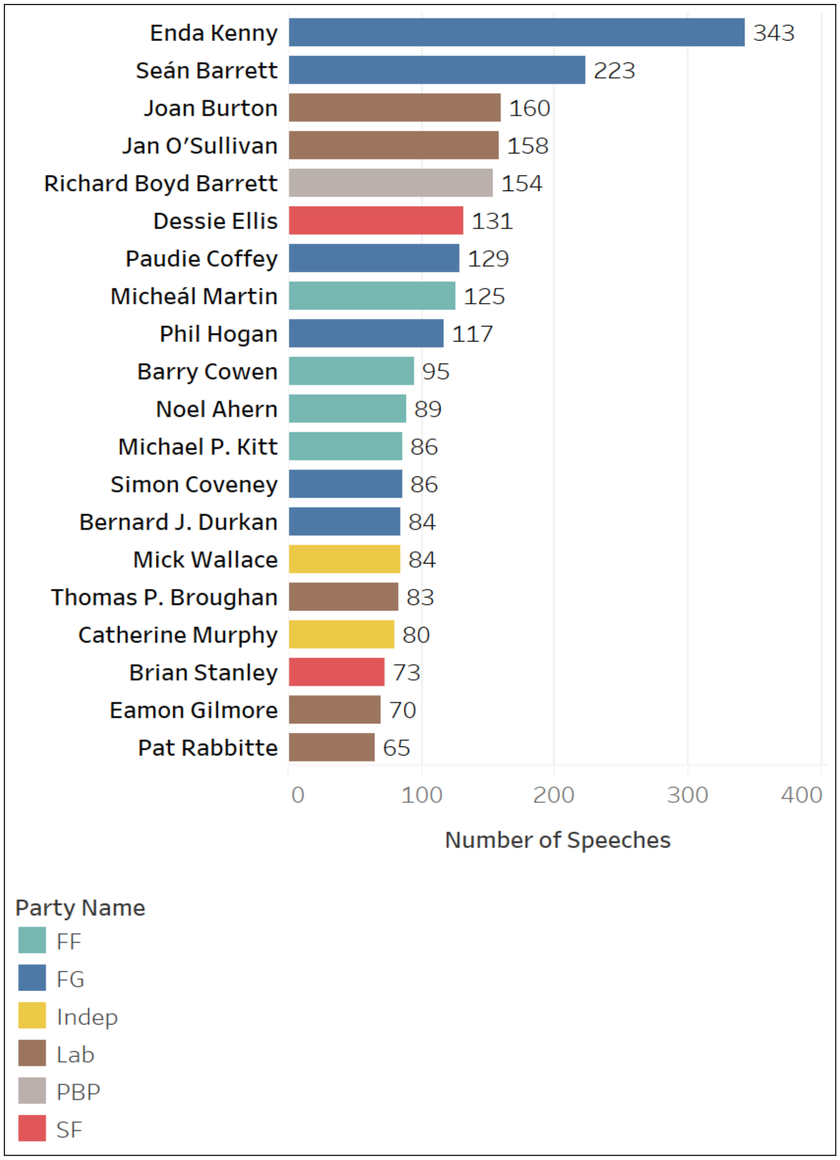


Figure 4.19: TDs by number of speeches (Top 20)

4.5 Further Discussion

The number of lead speeches versus follow-up speeches were looked at across all topics. However, a manual review of some samples suggests that the TD who posed the question is not always the lead speaker. For this reason, this could not be relied on to make any observations as to whom is asking the most questions. It does, however, suggest an opportunity for future work if this information could be retrieved.

The goal of this project is to get an insight into the activity of the Dáil and in particular what are the priorities of the TDs. The best way to view this is by looking at the spread of speeches for each individual TD across all topics. This will highlight what areas TDs have prioritised to discuss in the Dáil. Conversely, it will also uncover what topics they have chosen not to speak about. Is this because they have no interest in these areas or know nothing about them? Figures 4.23 and 4.24 (page 77) are heat maps which represent the percentage of speeches that a TD has made across all 18 topic areas for the independents and socialists respectively. The heatmap tables for the remaining parties been included in the appendix for reference. Take for example figure 4.23 which represents the independents. An individual row in this table represents the percentage of speeches that a TD has made on each topic out of the total speeches a TD has made. Looking vertically at this table, it can be seen which TD is spending the most time discussing a particular topic. On review of these tables, the following findings are made:

1. Comparing all parties together, there are some interesting findings by looking at the percentage of TDs within a party that have covered all 18 topics, as illustrated in figure 4.20. People before Profit, having 2 active speakers in the Dáil (Richard Boyd Barrett and Joan Collins), have covered all 18 topics. Sinn Féin come in with a remarkable 93% of their TDs speaking about all 18 topics. Next, is Fine Gael with 55% coverage followed by Labour and Socialists both with 50% coverage. Closely followed is the Independents at 43%. Finally, at the lower end of the scale is Fianna Fáil at 18% and the Green Party at 0%. This is unusual given that the Fianna Fáil party held office for over 50% of the period being analysed.

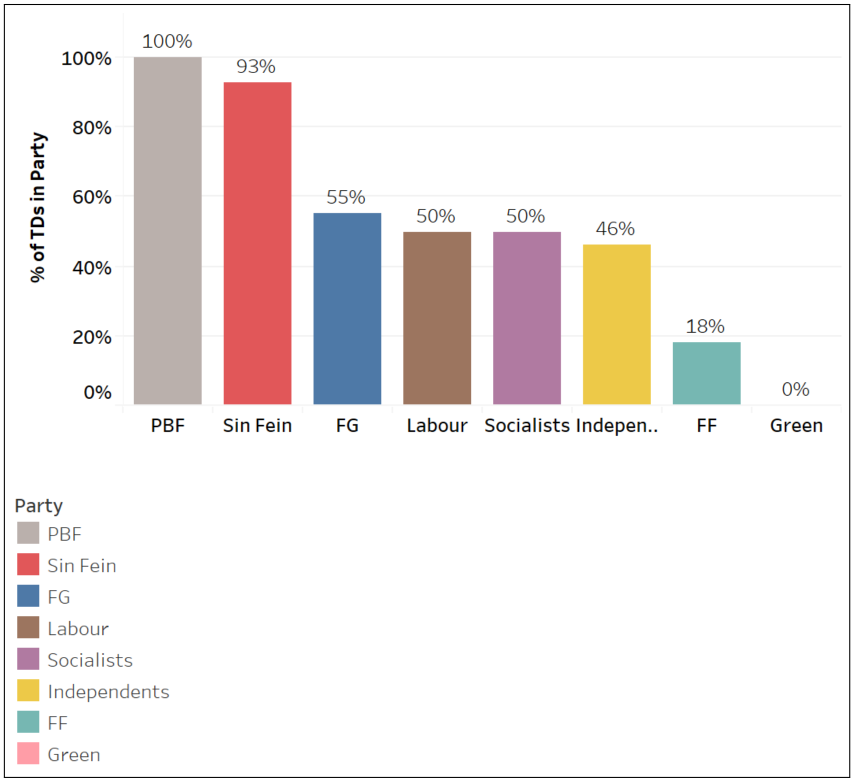


Figure 4.20: % of TDs in a party which have covered all 18 topics

2. Figure 4.21 illustrates the TDs per party which have 50% or less coverage on all topics which means that they have covered 9 or less topics out of the 18 generated. To provide some additional context, the number of years is also depicted in this graph which represents the number of years the TD was making speeches in the Dáil. It does not reflect their actual number of years in office but has the potential to explain the first finding further. There are TDs which have a low coverage on topics which may be accounted for the fact that they were only speaking during a short period of time in the analysis period such examples are John Dennehy and John Ellis. It also highlights some outliers such as Christy O'Sullivan whom has little over 30% coverage of topics but was speaking for 50% of the analysis period.

3. Fianna Fáil have 19 TDs which have less than 50% coverage on topics which in comparison to Fine Gael, a party of an equivalent size, is an interesting finding. However, drilling into this further 16 of these TDs were only making speeches for 20% or less of the 10-year analysis period. TDs such as Joe Walsh, whose main

priority was speaking about social care issues, only served as a sitting Dáil member for one-year of the 10-year analysis period which is reflected in the years he made speeches in figure 4.21. This would explain poor coverage across all topics for similar TDs.

4. Fine Gael's priorities are focused on more general/order of the day matters, most TDs have covered most topics at some point in time. There are few exceptions to this such as Gabrielle McFadden, Gay Mitchell and George Lee but all of these TDs spoke only for 20% of the analysis period. However, it is a small number in comparison to Fianna Fáil.

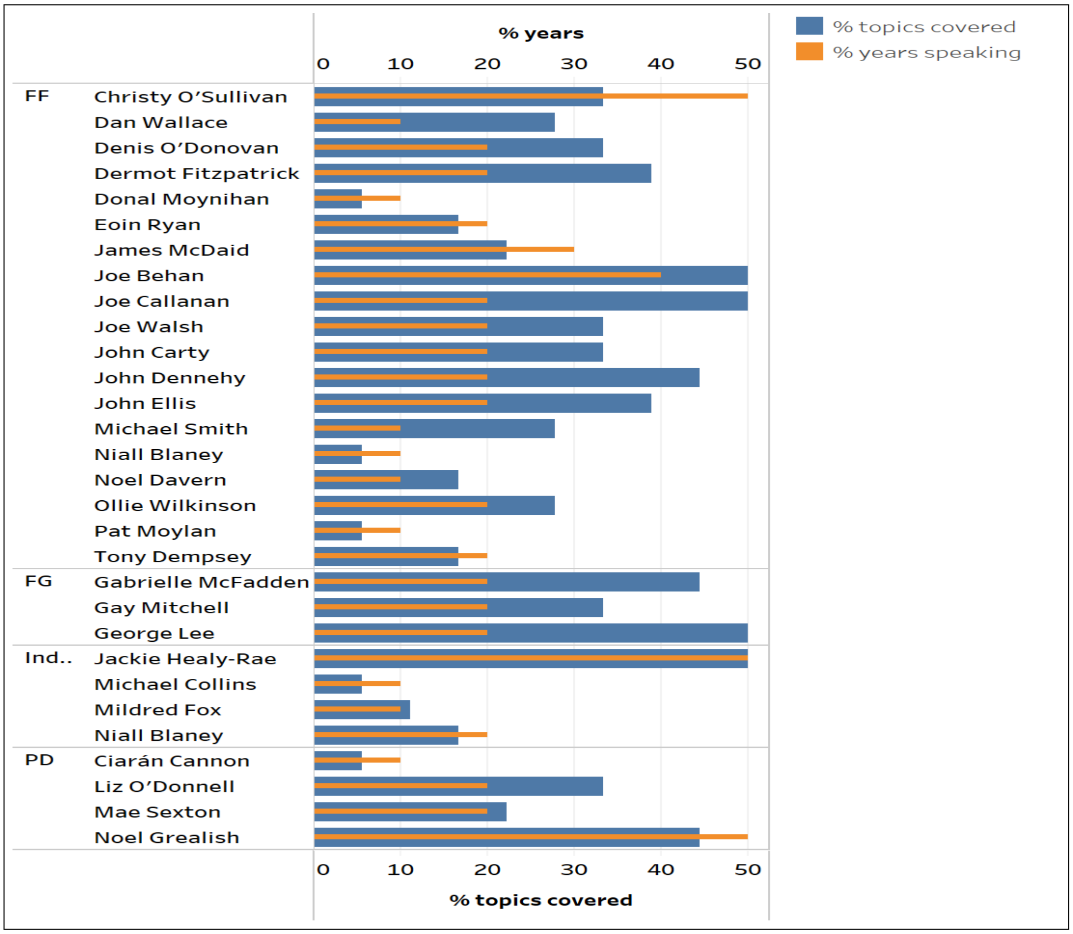


Figure 4.21: TDs - less than 50% topic coverage - % topic coverage versus % years speaking

5. Finally, figure 4.22 on page 76, drills into the Fianna Fáil party in a little more

detail providing a summary of the number of topics covered versus number of years a TD was speaking for. It also illustrates the average number of topics (14 out of 18) for this party to identify those TDs which are above and below this reference point, whilst still considering their number of years making speeches. It can be seen that there are a large number of TDs with a coverage over the average and some of these TDs are only speaking for 5 or less years of the analysis period such as Robert Troy, Barry Cowen, Charlie McConalogue, Mary Hanafin, Martin Cullen and Brian Lenihan Junior.

4.6 Summary of Analysis

This chapter has provided the results of each phase of the experiment. The experiment resulted in topics being generated with acceptable levels of coherence. A detailed analysis was conducted which manually labelled the topics and drilled into the detail of the underlying speeches to confirm the quality of them. It was found that both general and specific topics were generated and samples were taken of the niche topics. Finally, the trends in the topics verus external factors at the time provide further confirmation of the quality of the clusters generated. The main trends will be summarised in the conclusion chapter.

The next chapter concludes with a final summary of this study and suggests areas for future research.

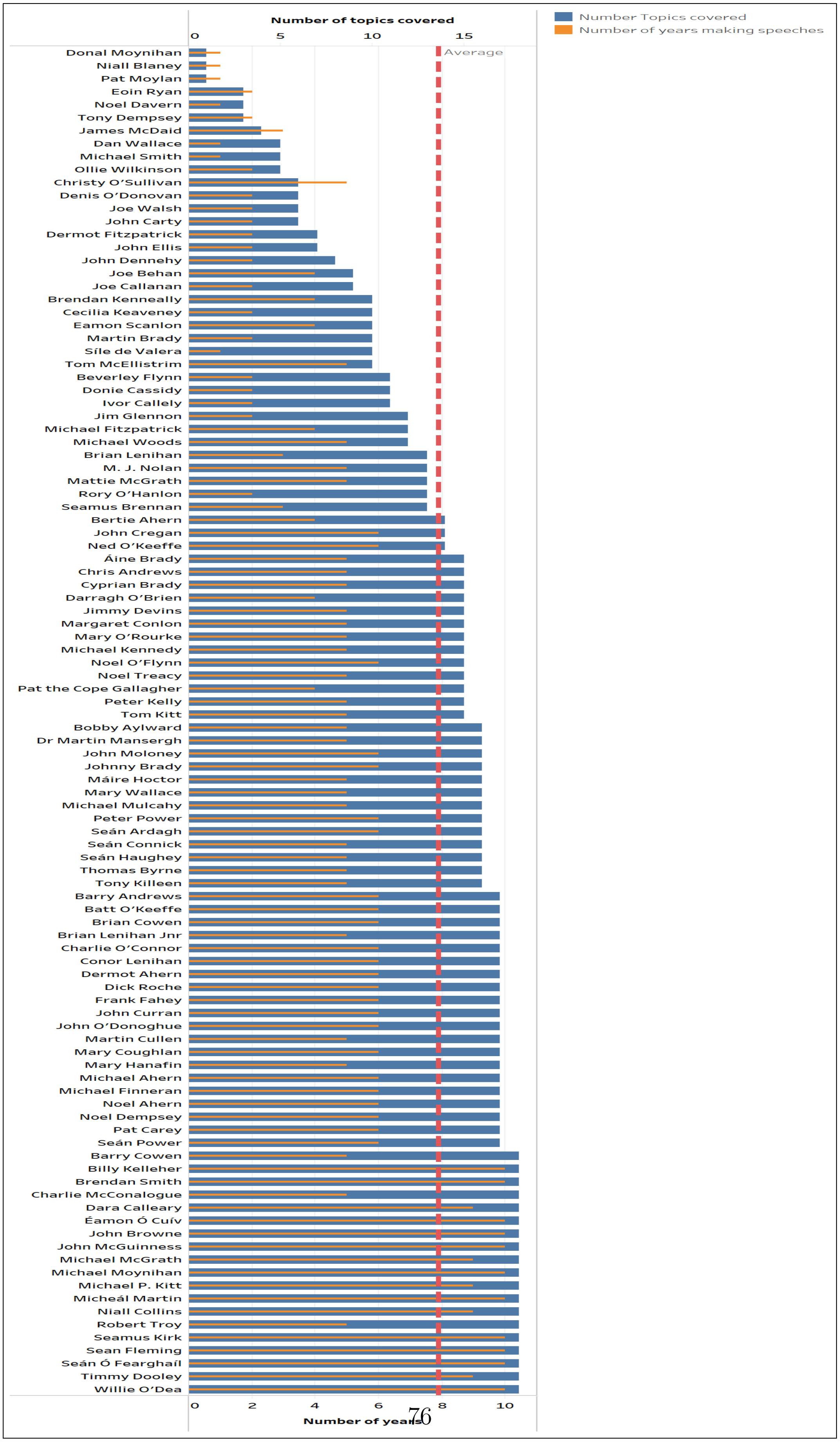
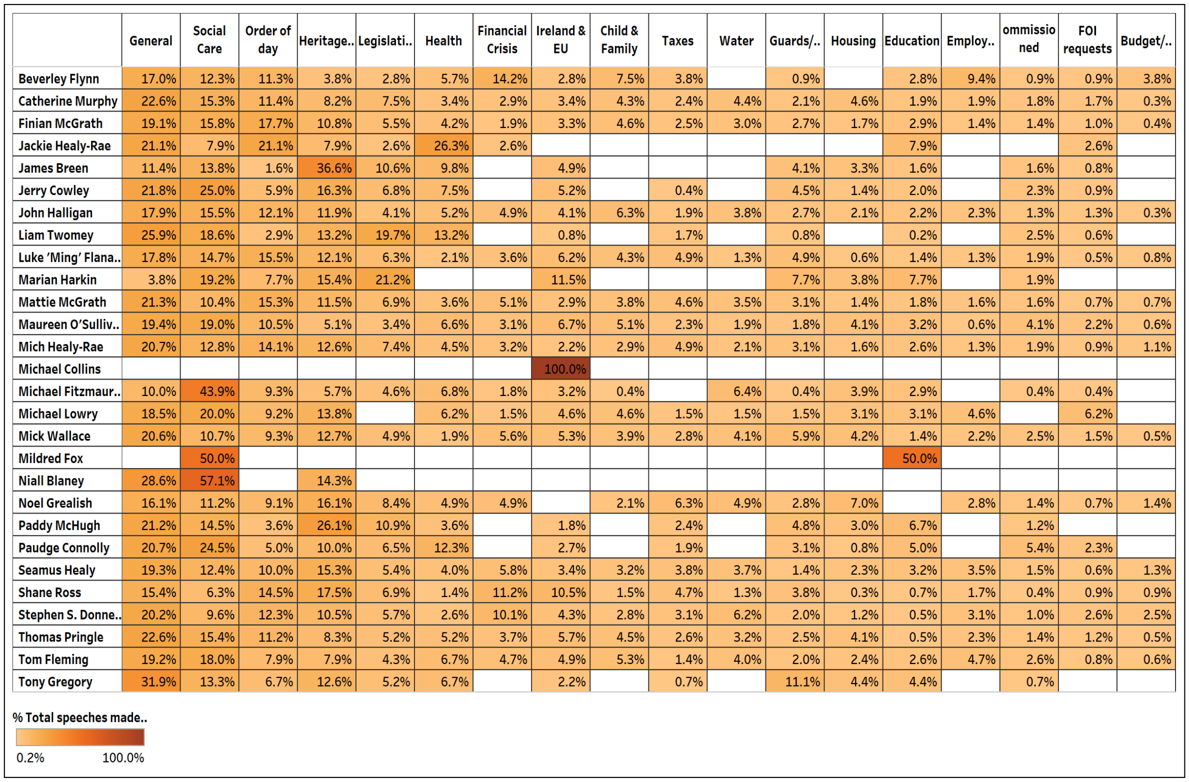


Figure 4.22: Fianna Fáil TDs - number topics versus number years speaking

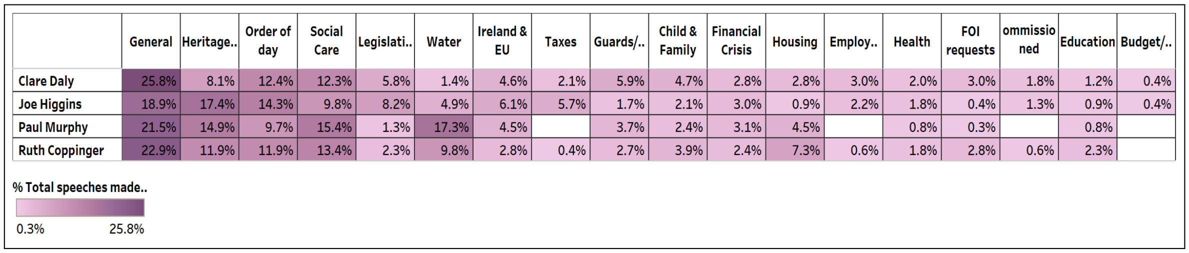


	General	Social Care	Order of day	Heritage..	Legislati..	Health	Financial Crisis	Ireland & EU	Child & Family	Taxes	Water	Guards/..	Housing	Education	Employ..	ommissioned	FOI requests	Budget/..
Beverley Flynn	17.0%	12.3%	11.3%	3.8%	2.8%	5.7%	14.2%	2.8%	7.5%	3.8%		0.9%		2.8%	9.4%	0.9%	0.9%	3.8%
Catherine Murphy	22.6%	15.3%	11.4%	8.2%	7.5%	3.4%	2.9%	3.4%	4.3%	2.4%	4.4%	2.1%	4.6%	1.9%	1.9%	1.8%	1.7%	0.3%
Finian McGrath	19.1%	15.8%	17.7%	10.8%	5.5%	4.2%	1.9%	3.3%	4.6%	2.5%	3.0%	2.7%	1.7%	2.9%	1.4%	1.4%	1.0%	0.4%
Jackie Healy-Rae	21.1%	7.9%	21.1%	7.9%	2.6%	26.3%	2.6%							7.9%			2.6%	
James Breen	11.4%	13.8%	1.6%	36.6%	10.6%	9.8%		4.9%				4.1%	3.3%	1.6%		1.6%	0.8%	
Jerry Cowley	21.8%	25.0%	5.9%	16.3%	6.8%	7.5%		5.2%		0.4%		4.5%	1.4%	2.0%		2.3%	0.9%	
John Halligan	17.9%	15.5%	12.1%	11.9%	4.1%	5.2%	4.9%	4.1%	6.3%	1.9%	3.8%	2.7%	2.1%	2.2%	2.3%	1.3%	1.3%	0.3%
Liam Twomey	25.9%	18.6%	2.9%	13.2%	19.7%	13.2%		0.8%		1.7%		0.8%		0.2%		2.5%	0.6%	
Luke 'Ming' Flana..	17.8%	14.7%	15.5%	12.1%	6.3%	2.1%	3.6%	6.2%	4.3%	4.9%	1.3%	4.9%	0.6%	1.4%	1.3%	1.9%	0.5%	0.8%
Marian Harkin	3.8%	19.2%	7.7%	15.4%	21.2%			11.5%				7.7%	3.8%	7.7%		1.9%		
Mattie McGrath	21.3%	10.4%	15.3%	11.5%	6.9%	3.6%	5.1%	2.9%	3.8%	4.6%	3.5%	3.1%	1.4%	1.8%	1.6%	1.6%	0.7%	0.7%
Maureen O'Sulliv..	19.4%	19.0%	10.5%	5.1%	3.4%	6.6%	3.1%	6.7%	5.1%	2.3%	1.9%	1.8%	4.1%	3.2%	0.6%	4.1%	2.2%	0.6%
Mich Healy-Rae	20.7%	12.8%	14.1%	12.6%	7.4%	4.5%	3.2%	2.2%	2.9%	4.9%	2.1%	3.1%	1.6%	2.6%	1.3%	1.9%	0.9%	1.1%
Michael Collins								100.0%										
Michael Fitzmaur..	10.0%	43.9%	9.3%	5.7%	4.6%	6.8%	1.8%	3.2%	0.4%		6.4%	0.4%	3.9%	2.9%		0.4%	0.4%	
Michael Lowry	18.5%	20.0%	9.2%	13.8%		6.2%	1.5%	4.6%	4.6%	1.5%	1.5%	1.5%	3.1%	3.1%	4.6%		6.2%	
Mick Wallace	20.6%	10.7%	9.3%	12.7%	4.9%	1.9%	5.6%	5.3%	3.9%	2.8%	4.1%	5.9%	4.2%	1.4%	2.2%	2.5%	1.5%	0.5%
Mildred Fox		50.0%												50.0%				
Niall Blaney	28.6%	57.1%		14.3%														
Noel Grealish	16.1%	11.2%	9.1%	16.1%	8.4%	4.9%	4.9%		2.1%	6.3%	4.9%	2.8%	7.0%		2.8%	1.4%	0.7%	1.4%
Paddy McHugh	21.2%	14.5%	3.6%	26.1%	10.9%	3.6%		1.8%		2.4%		4.8%	3.0%	6.7%		1.2%		
Paudge Connolly	20.7%	24.5%	5.0%	10.0%	6.5%	12.3%		2.7%		1.9%		3.1%	0.8%	5.0%		5.4%	2.3%	
Seamus Healy	19.3%	12.4%	10.0%	15.3%	5.4%	4.0%	5.8%	3.4%	3.2%	3.8%	3.7%	1.4%	2.3%	3.2%	3.5%	1.5%	0.6%	1.3%
Shane Ross	15.4%	6.3%	14.5%	17.5%	6.9%	1.4%	11.2%	10.5%	1.5%	4.7%	1.3%	3.8%	0.3%	0.7%	1.7%	0.4%	0.9%	0.9%
Stephen S. Donne..	20.2%	9.6%	12.3%	10.5%	5.7%	2.6%	10.1%	4.3%	2.8%	3.1%	6.2%	2.0%	1.2%	0.5%	3.1%	1.0%	2.6%	2.5%
Thomas Pringle	22.6%	15.4%	11.2%	8.3%	5.2%	5.2%	3.7%	5.7%	4.5%	2.6%	3.2%	2.5%	4.1%	0.5%	2.3%	1.4%	1.2%	0.5%
Tom Fleming	19.2%	18.0%	7.9%	7.9%	4.3%	6.7%	4.7%	4.9%	5.3%	1.4%	4.0%	2.0%	2.4%	2.6%	4.7%	2.6%	0.8%	0.6%
Tony Gregory	31.9%	13.3%	6.7%	12.6%	5.2%	6.7%		2.2%		0.7%		11.1%	4.4%	4.4%		0.7%		

% Total speeches made..

0.2% 100.0%

Figure 4.23: Independents -TD by topic coverage % of total speeches made by a TD



	General	Heritage..	Order of day	Social Care	Legislati..	Water	Ireland & EU	Taxes	Guards/..	Child & Family	Financial Crisis	Housing	Employ..	Health	FOI requests	ommissioned	Education	Budget/..
Clare Daly	25.8%	8.1%	12.4%	12.3%	5.8%	1.4%	4.6%	2.1%	5.9%	4.7%	2.8%	2.8%	3.0%	2.0%	3.0%	1.8%	1.2%	0.4%
Joe Higgins	18.9%	17.4%	14.3%	9.8%	8.2%	4.9%	6.1%	5.7%	1.7%	2.1%	3.0%	0.9%	2.2%	1.8%	0.4%	1.3%	0.9%	0.4%
Paul Murphy	21.5%	14.9%	9.7%	15.4%	1.3%	17.3%	4.5%		3.7%	2.4%	3.1%	4.5%		0.8%	0.3%		0.8%	
Ruth Coppinger	22.9%	11.9%	11.9%	13.4%	2.3%	9.8%	2.8%	0.4%	2.7%	3.9%	2.4%	7.3%	0.6%	1.8%	2.8%	0.6%	2.3%	

% Total speeches made..

0.3% 25.8%

Figure 4.24: Socialists -TD by topic coverage by % of total speeches made by a TD

Chapter 5

Conclusion

5.1 Research Overview

This study analysed the verbal parliamentary speeches in the Dáil over a ten-year period. The process commenced by conducting an in-depth literature review of text mining techniques and in particular those used to classify or cluster political textual data. A comparison was made between supervised and unsupervised machine learning with examples of techniques across both. The main area of this review then focused on unsupervised machine learning techniques as the available data set was unlabelled.

Following this, as previous research demonstrated that topic modeling was successful in clustering political speeches in the European Parliament, this area was investigated further. For this reason, two topic modelling techniques were reviewed. The first of those was Latent Dirichlet Allocation (LDA) followed by Non-negative Matrix Factorisation (NMF). LDA treats textual data as a combination of probability distributions. Each word in the document and the document itself can both be represented as a distribution over a set of clusters/topics. NMF uses a matrix decomposition technique to reduce the data into two factors that are constrained so as not to contain any negative values. Each non-negative factor can be represented as the additive combination of a set of non-negative basis vectors which in turn produce a clustering of the data into topics. Both of these techniques have been used in the clustering of documents in varying domains but NMF has shown to provide more coherent topics when being applied to niche subject matter domains such as the European parliamentary data set. For this reason, NMF is the algorithm chosen to apply to the Irish parliamentary speeches. In order to see how the themes evolved over time, a two-layer dynamic topic modeling approach was used to extract themes at a point in time and over the entire period.

5.2 Problem Definition

This study utilises an existing unsupervised machine learning technique but applies it to a data set which has never been analysed using such techniques. More specifically, the main goal of this work is to explore the application of NMF on the Irish parliamentary questions to produce coherent themes so that they can in turn be analysed to highlight trends in parliamentary activity. This work utilised both quantitative and qualitative analysis to compile the final results and perform an evaluation of them. The research question that this thesis chose to examine is as follows:

"Can a two-layer NMF dynamic topic model yield coherent topics in order to highlight trends in the Irish Parliamentary speeches data set?"

5.3 Design/Experimentation, Evaluation & Results

The key stages in this study can be summarised as follows:

1. Pre-processing.

2. Modeling and refinement.

3. Evaluation.

4. Manual review of topics and identify appropriate labels.

5. Visualisation of final results and trends.

The very first step involved reviewing the available data set to assess the quality of it. At the early stage, it was clear that there were data quality issues that had to be dealt with. Data cleansing involved having to retrieve some of the data again from the Oireachtas website due to a parsing error in the speech content. Also, cleansing involved replacing erroneous characters with their correct equivalents or deleting them from the content. An additional data set was retrieved to ensure that TDs were represented more accurately according to what parties they were a member of. This was mainly to account for the fact that TDs could move between parties as the original data set did not account for this. The final data set collected was a list of parties that were in government across the Dáil terms being analysed. This provided further information on whether a TD was a member of the opposition or government at the time they made a particular speech. Considerably effort was spent on merging these data sets correctly, some of which required manual effort.

Following data cleaning tasks, the speeches data set was partitioned into 40 quarterly time windows. Then the speech content was processed by stripping out stop words and words in a minimum and maximum number of documents. Nouns and proper nouns were extracted only and plural nouns were replaced with their singular version. The optimal pre-processing conditions were tested using a single time window and a selected static value for the number of topics. Then, the appropriate number of topics for all 40 time windows was selected using coherence measured against their term representations in a word2vec model space. This produced 40 window topic models with 494 static window topics. These window topic documents were then used as the input to the second dynamic layer of the NMF model. Using the same methodology as before, the appropriate number of dynamic topics was selected. The final dynamic model produced 26 topics which relate back to 494 window topics and their underlying speech documents. Coherence was used to evaluate the final model in a mathematical manner. Alongside this, the manual labelling of the final dynamic and window topics is the key evidence that the experiment successfully extracted themes from the underlying speeches. To complement this further, detailed analysis was carried out on examples of dynamic topics generated to confirm their quality by reviewing the top ranked terms, their top ranked speeches and finally against external factors.

The final topics generated range from 16 niche areas to 10 more general matters. It can be seen how these themes evolved over time by their spread across quarters. The contribution of these topics by TD and party illustrate what they have spoken about over this period and provide insight into their political priorities. Conversely, by reviewing the topics that TDs have not discussed gives an insight into the themes that they have no interest in. Ignoring the general topics, the most widely discussed issues are social care, heritage/environment and legislation which have been discussed in more than 33 quarters over the ten-year period. On the opposite end of the scale, budget/finance, housing and water have the least number of speeches and are discussed in less than 9 quarters. The top three holds across all parties, however, the bottom three only holds true for the larger parties namely Fine Gael, Fianna Fáil and Labour. The socialist party and people before profit have illustrated a higher preference for discussing water and housing issues in comparison. Sinn Féin have given less priority to discussing FOI requests and education. The independents also illustrate less preference towards FOI requests and employment matters. The smallest contribution of speeches to any one topic was made by the Green party in relation to water issues, at zero speeches, followed by People Before Profit with 36 speeches relating to budget/finance issues.

5.4 Contributions and Impact

The detailed exploration of the data set documented in this research has the potential to be of benefit to future researchers as follows:

- First and foremost, it is the first time that unsupervised learning has been applied to the Irish parliamentary speeches data set. Therefore, this will inform political scientists in Ireland and abroad that such an approach may be used. This is valuable knowledge given the increasing volumes of unclassified and unstructured political text and documents now being made available online.

- Secondly, it is using an unsupervised machine learning algorithm (NMF) which is under-utilised in research to date in comparison to other algorithms such as LDA. Therefore, it is a valuable addition to the computer science knowledge domain.
- Finally, it illustrates to interested parties what TDs and political parties have been prioritising as part of verbal question time and more interestingly what they have not discussed. The study reveals what percentage of TDs in each party have spoken on all topics with the Green party and Fianna Fáil at the lower end of the scale.

5.5 Future Work & Recommendations

An important aspect to this work is that it is related to the verbal parliamentary speeches only so whilst the results are indicative of political priorities, it is not the complete picture. As stated in the literature review chapter, TDs ask a lot of detailed questions in written format. Future work could involve analysing the written parliamentary questions made available online in conjunction with the verbal speeches. The combination of the written and verbal questions would provide a more comprehensive data set and combined with the approach used in this study has the potential to provide an over-arching tool for monitoring the activities and priorities of the Irish Parliament.

Another key area worth investigating further are the gerenal/order of the day topics to see if more meaningful themes can be extracted from them. Finally, some additional coherence measures could be used to substantiate the semantic validity of the topics further such as TC-NPMI (Normalised PMI) and TC-LCP (Mean pairwise log conditional probability) as utilised by OCallaghan et al. (2015).

As stated in the literature review, a large number of resources are tied up with answering PQs in State departments and Agencies which the taxpayer is paying for. For this reason, it is important to understand the themes being raised and the quality of the questions being posed. This will only become apparent when there is an adequate means of tracking and monitoring these activitites. Some addtional information that would be useful in this type of analysis would be as follows:

- Which TD posed the question?
- Which state department or agency is tied up with providing the answer?
- How much money has it cost the taxpayer in providing the final response?

The above additional data may force TDs to ask more well-thought-out and meaningful questions as opposed to it being a numbers game.

References

Ahmadi, P., Tabandeh, M., & Gholampour, I. (2016). Persian text classification based on topic models. In *Electrical Engineering (ICEE), 2016 24th Iranian Conference on* (pp. 86–91). IEEE.

Arora, S., Ge, R., & Moitra, A. (2012, October). Learning Topic Models Going beyond SVD. In *2012 IEEE 53rd Annual Symposium on Foundations of Computer Science* (pp. 1–10).

Basu, A., Walters, C., & Shepherd, M. (2003). Support vector machines for text categorization. In *System Sciences, 2003. Proceedings of the 36th Annual Hawaii International Conference on* (pp. 7–pp). IEEE.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77–84.

Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning* (pp. 113–120). ACM.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993–1022.

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory* (pp. 144–152). ACM.

Boutsidis, C., & Gallopoulos, E. (2008, April). SVD based initialization: A head start for nonnegative matrix factorization. *Pattern Recognition*, *41*(4), 1350–1362.

Bulut, A. T. (2016, January). Measuring political agenda setting and representation in Turkey: Introducing a new approach and data set. *Party Politics*.

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems* (pp. 288–296).

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). Crisp-dm 1.0 step-by-step data mining guide.

Dandoy, R. (2011, September). Parliamentary Questions in Belgium: Testing for Party Discipline. *The Journal of Legislative Studies*, *17*(3), 315–326.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, *41*(6), 391.

Delany, S. J., Sinnott, R., & O'Reilly, N. (2010). The extent of clientelism in Irish politics: Evidence from classifying Dil questions on a local-national dimension. In *AICS: Proceedings of 21st Irish Conference on Artificial Intelligence and Cognitive Science, NUI Galway, 30 August - 1st September.*

Greene, D., & Cross, J. P. (2015). Unveiling the Political Agenda of the European Parliament Plenary: A Topical Analysis. In *Proceedings of the acm web science conference* (pp. 1–10). ACM.

Greene, D., & Cross, J. P. (2017, January). Exploring the Political Agenda of the European Parliament Using a Dynamic Topic Modeling Approach. *Political Analysis*, *25*(01), 77–94.

Greene, D., OCallaghan, D., & Cunningham, P. (2014). How many topics? stability analysis for topic models. In T. Calders, F. Esposito, E. Hllermeier, & R. Meo (Eds.), *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (Vol. 8724, pp. 498–513). Springer Berlin Heidelberg.

Grimmer, J. (2010, January). A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases. *Political Analysis*, *18*(1), 1–35.

Laitonjam, N., Padmanabhan, V., Pujari, A. K., & Lal, R. P. (2015, December). Topic Modelling for Songs. In (pp. 130–135). IEEE.

Lee, D. D., & Seung, H. S. (1999, October). Learning the parts of objects by non-negative matrix factorization. *Nature*, *401*, 788–91.

Lee, H., Hong, B., & Kim, K. K. (2015). Documents topic classification model in social networks using classifiers voting system. In (pp. 68–73). ACM.

Lee, S., Kim, J., & Myaeng, S.-H. (2015). An extension of topic models for text classification: A term weighting approach. In *2015 International Conference on Big Data and Smart Computing (BIGCOMP)* (pp. 217–224). IEEE.

Li, Z., Shang, W., & Yan, M. (2016). News text classification model based on topic model. In *Computer and Information Science (ICIS), 2016 IEEE/ACIS 15th International Conference on* (pp. 1–5). IEEE.

Lin, C.-J. (2007). Projected gradient methods for nonnegative matrix factorization. *Neural computation*, *19*(10), 2756–2779.

Lu, B., Ott, M., Cardie, C., & Tsou, B. K. (2011, December). Multi-aspect Sentiment Analysis with Topic Models. In (pp. 81–88). IEEE.

Martin, S. (2011, June). Using Parliamentary Questions to Measure Constituency Focus: An Application to the Irish Case. *Political Studies*, *59*(2), 472–488.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mimno, D., Wallach, H. M., Talley, E., Leenders, M., & McCallum, A. (2011). Optimizing semantic coherence in topic models. In *Proceedings of the conference* *on empirical methods in natural language processing* (pp. 262–272). Association for Computational Linguistics.

Newman, D., Lau, J. H., Grieser, K., & Baldwin, T. (2010). Automatic evaluation of topic coherence. In *Human language technologies: The 2010 annual conference of the north american chapter of the association for computational linguistics* (pp. 100–108). Stroudsburg, PA, USA: Association for Computational Linguistics.

OCallaghan, D., Greene, D., Carthy, J., & Cunningham, P. (2015). An analysis of the coherence of descriptors in topic modeling. *Expert Systems with Applications*, *42*(13), 5645–5657.

Papadimitriou, C. H., Tamaki, H., Raghavan, P., & Vempala, S. (1998). Latent semantic indexing: A probabilistic analysis. In *Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems* (pp. 159–168). ACM.

Quinn, K. M., Monroe, B. L., Colaresi, M., Crespin, M. H., & Radev, D. R. (2010). How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science*, *54*(1), 209–228.

Roder, M., Both, A., & Hinneburg, A. (2015). Exploring the Space of Topic Coherence Measures. In (pp. 399–408). ACM.

Rozenberg, O., & Martin, S. (2011, September). Questioning Parliamentary Questions. *The Journal of Legislative Studies*, *17*(3), 394–404.

Salmond, R. (2014, July). Parliamentary Question Times: How Legislative Accountability Mechanisms Affect Mass Political Engagement. *The Journal of Legislative Studies*, *20*(3), 321–341.

Stevens, K., Kegelmeyer, P., Andrzejewski, D., & Buttler, D. (2012). Exploring topic coherence over many models and many topics. In *Proceedings of the 2012*

Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (pp. 952–961). Association for Computational Linguistics.

Steyvers, M., & Griffiths, T. (2007). Latent semantic analysis: A road to meaning. In T. Landauer, S. D. McNamara, & W. Kintsch (Eds.), (chap. Probabilistic topic models). Laurence Erlbaum.

Vliegenthart, R., Walgrave, S., Baumgartner, F. R., Bevan, S., Breunig, C., Brouard, S., . . . Tresch, A. (2016, May). Do the media set the parliamentary agenda? A comparative study in seven countries. *European Journal of Political Research*, *55*(2), 283–301.

Wang, J. J.-Y., Wang, X., & Gao, X. (2013). Non-negative matrix factorization by maximizing correntropy for cancer clustering. *BMC bioinformatics*, *14*(1), 107.

Wu, J. (2012). *Advances in k-means clustering* (1st ed.). New York City: Springer-Verlag Berlin Heidelberg.

Yu, B., Kaufmann, S., & Diermeier, D. (2008, July). Classifying Party Affiliation from Political Speech. *Journal of Information Technology & Politics*, *5*(1), 33–48.

Appendix A

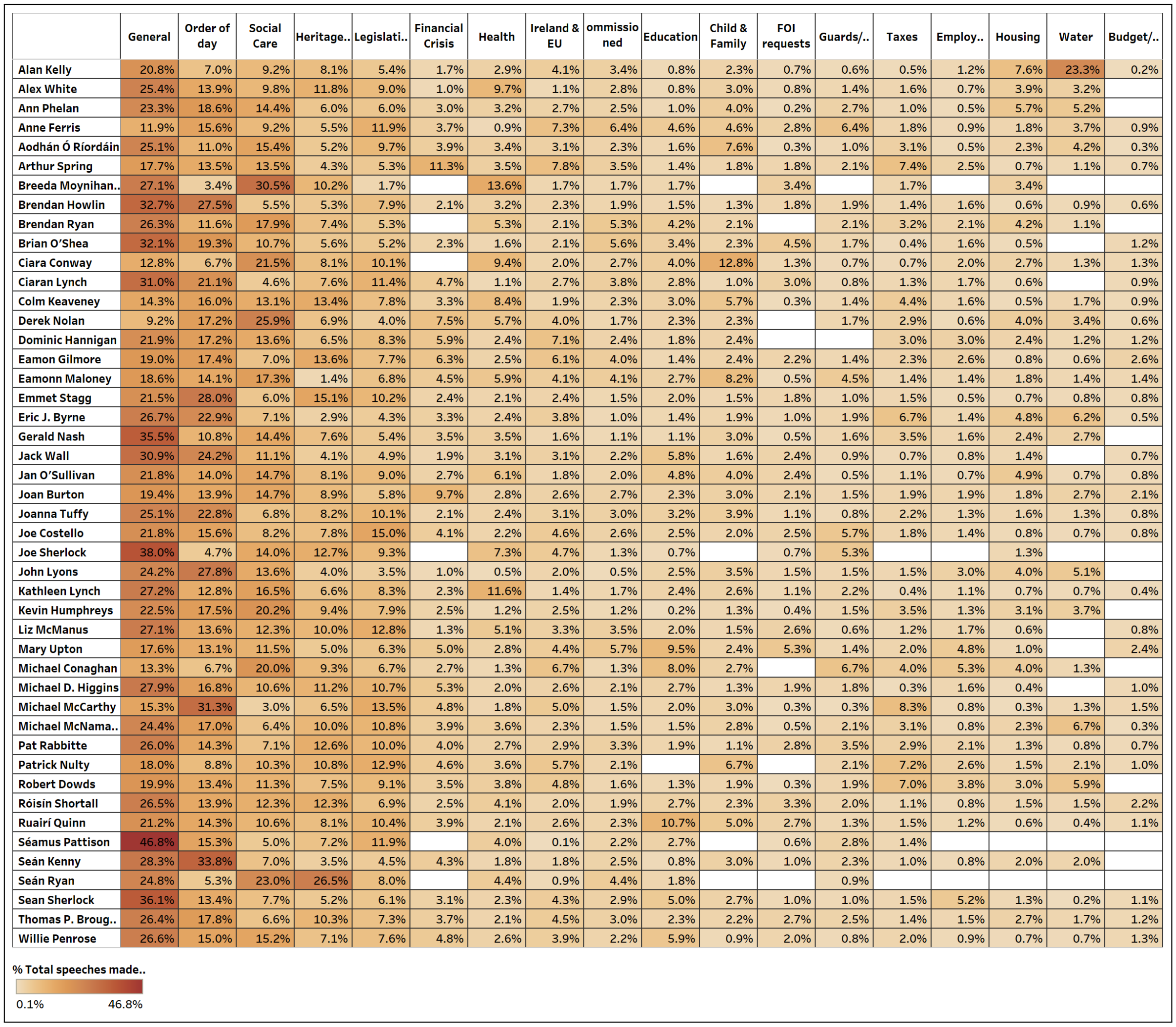
Additional tables and figures

A.1 Top ranked speeches in Financial Crisis topic across all time windows

A.2 Heat Maps - Topics and TDs by % of total speeches a TD made

Speech ID	Dail	TD	Speech Excerpt
123104	30	John O'Donoghue	Asking on the Order of Business who pays if a bank goes under -come on Deputy Gilmore.
139602	30	Brian Cowen	The Anglo Irish Bank which is the bank to which the Deputy referred.
170628	30	Michael Ahern	In the 1970s I worked as a financial controller in a large building firm in Cork. At that time the country was in the throes of an economic crisis not much different and possibly worse than that being experienced today. I know first hand the problems..
206085	30	Michael D. Higgins	Anglo Irish Bank the systemic bank.
208512	30	Sean Fleming	I welcome the opportunity to speak on the Central Bank Reform Bill 2010. I want to deal with the specifics of the legislation and then outline the further measures that will be needed to regulate Irish banks beyond the contents of the legislation. After..
292651	31	Enda Kenny	We called in the banks and had a discussion with them about lending policy in general and whether AIB and Bank of Ireland would be in a position to meet the 2011 target of lending 3 billion apiece 3.5 billion next year and 4 billion in 2013. We...
113997	30	Seymour Crawford	What about Ulster Bank?
151720	30	Joan Burton	..and banks not covered. The Tanaiste needs to enlighten us..
179405	30	Padraic McCormack	Is that the banks statement?
226920	30	Kieran O'Donnell	....The guarantee reminds me of the original guarantee scheme which was introduced on 30 September 2008. We were not provided with the full facts then. I remember on the night that the question was asked why the Government was including dated subordinated debt lower tier 2 but there was no answer. We still do not know the final cost of Anglo Irish Bank. Why was the cost not announced prior to the extension of the ELG scheme? The cost of borrowing..
231668	30	Brendan Howlin	Perhaps the Deputy could bank them.
304140	31	Michael Martin	Is the Taoiseach saying the bank is not relevant?
328647	31	Richard Boyd Barrett	The question is who the banks serve.
339566	31	Michael McGrath	.....Credit unions like all other financial institutions in this State have been affected by the recession and downturn in our economy since 2008. It must acknowledged however that the scale of the problems in credit unions while significant is in no way comparable to those which have emerged in our banking system. Credit unions like all other financial institutions in this State have been affected by the recession and downturn in our economy since 2008. It must acknowledged however that the scale of the problems in credit unions while significant is in no way comparable to those which have emerged in our banking system.
362839	31	John Halligan	The bank are trying to take her house from her. I am upset because of the cases I am hearing about every day.
366955	31	Mattie McGrath	Do not get me stuck on the banks issue. Deputy Arthur Spring was there himself.
382989	31	Michael McGrath	...As I mentioned the evidence given by the banks two weeks ago took me by surprise and it took every Government Deputy present by surprise as well. It probably even took the Government by surprise. To be fair to the Government and the Central Bank - I will be critical of both in a moment - when the targets were issued last March the last thing both the Government and Central Bank expected was that up to 15 000 letters would be issued by banks to those in mortgage arrears.....
399299	31	Michael Martin	They gave the power to the banks.
424525	31	Enda Kenny	We have figures from the Central Bank the Department of Finance and the banks themselves.
458298	31	Michelle Mulherin	I welcome the opportunity to speak on the matters raised in this Bill. As long as many of our citizens struggle with mortgage debt,we have to keep revisiting this issue. The banks have been given targets and timelines for restructuring and dealing with individuals in arrears issue. The banks have been given targets and timelines for restructuring and dealing with individuals in arrears..
477141	31	Sean Fleming	...This is outstanding careful and measured legislation. The crunch issue is the proposal to require the Central Bank to carry out an assessment of the state of the mortgage market. Should the Central Bank conclude that a market failure exists the legislation provides that it would be empowered with a range of tools to influence standard variable interest rates..

Table A.1: Financial Crisis - top ranked speech in 21 quarters that topic appears in

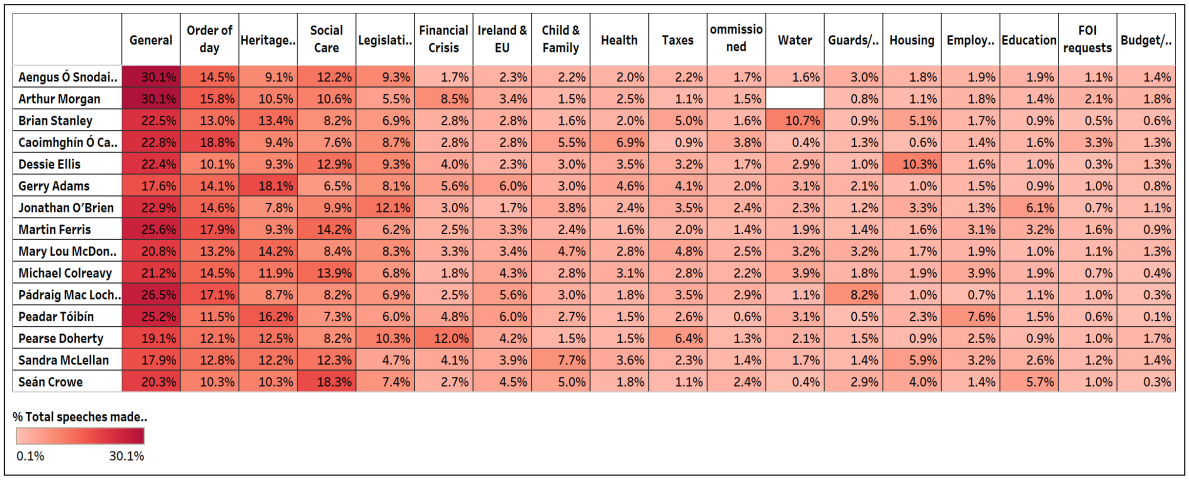


	General	Order of day	Social Care	Heritage..	Legislati..	Financial Crisis	Health	Ireland & EU	ommissio ned	Education	Child & Family	FOI requests	Guards/..	Taxes	Employ..	Housing	Water	Budget/..
Alan Kelly	20.8%	7.0%	9.2%	8.1%	5.4%	1.7%	2.9%	4.1%	3.4%	0.8%	2.3%	0.7%	0.6%	0.5%	1.2%	7.6%	23.3%	0.2%
Alex White	25.4%	13.9%	9.8%	11.8%	9.0%	1.0%	9.7%	1.1%	2.8%	0.8%	3.0%	0.8%	1.4%	1.6%	0.7%	3.9%	3.2%	
Ann Phelan	23.3%	18.6%	14.4%	6.0%	6.0%	3.0%	3.2%	2.7%	2.5%	1.0%	4.0%	0.2%	2.7%	1.0%	0.5%	5.7%	5.2%	
Anne Ferris	11.9%	15.6%	9.2%	5.5%	11.9%	3.7%	0.9%	7.3%	6.4%	4.6%	4.6%	2.8%	6.4%	1.8%	0.9%	1.8%	3.7%	0.9%
Aodhán Ó Ríordáin	25.1%	11.0%	15.4%	5.2%	9.7%	3.9%	3.4%	3.1%	2.3%	1.6%	7.6%	0.3%	1.0%	3.1%	0.5%	2.3%	4.2%	0.3%
Arthur Spring	17.7%	13.5%	13.5%	4.3%	5.3%	11.3%	3.5%	7.8%	3.5%	1.4%	1.8%	1.8%	2.1%	7.4%	2.5%	0.7%	1.1%	0.7%
Breeda Moynihan..	27.1%	3.4%	30.5%	10.2%	1.7%		13.6%	1.7%	1.7%	1.7%		3.4%		1.7%		3.4%		
Brendan Howlin	32.7%	27.5%	5.5%	5.3%	7.9%	2.1%	3.2%	2.3%	1.9%	1.5%	1.3%	1.8%	1.9%	1.4%	1.6%	0.6%	0.9%	0.6%
Brendan Ryan	26.3%	11.6%	17.9%	7.4%	5.3%		5.3%	2.1%	5.3%	4.2%	2.1%		2.1%	3.2%	2.1%	4.2%	1.1%	
Brian O'Shea	32.1%	19.3%	10.7%	5.6%	5.2%	2.3%	1.6%	2.1%	5.6%	3.4%	2.3%	4.5%	1.7%	0.4%	1.6%	0.5%		1.2%
Ciara Conway	12.8%	6.7%	21.5%	8.1%	10.1%		9.4%	2.0%	2.7%	4.0%	12.8%	1.3%	0.7%	0.7%	2.0%	2.7%	1.3%	1.3%
Ciaran Lynch	31.0%	21.1%	4.6%	7.6%	11.4%	4.7%	1.1%	2.7%	3.8%	2.8%	1.0%	3.0%	0.8%	1.3%	1.7%	0.6%		0.9%
Colm Keaveney	14.3%	16.0%	13.1%	13.4%	7.8%	3.3%	8.4%	1.9%	2.3%	3.0%	5.7%	0.3%	1.4%	4.4%	1.6%	0.5%	1.7%	0.9%
Derek Nolan	9.2%	17.2%	25.9%	6.9%	4.0%	7.5%	5.7%	4.0%	1.7%	2.3%	2.3%		1.7%	2.9%	0.6%	4.0%	3.4%	0.6%
Dominic Hannigan	21.9%	17.2%	13.6%	6.5%	8.3%	5.9%	2.4%	7.1%	2.4%	1.8%	2.4%			3.0%	3.0%	2.4%	1.2%	1.2%
Eamon Gilmore	19.0%	17.4%	7.0%	13.6%	7.7%	6.3%	2.5%	6.1%	4.0%	1.4%	2.4%	2.2%	1.4%	2.3%	2.6%	0.8%	0.6%	2.6%
Eamonn Maloney	18.6%	14.1%	17.3%	1.4%	6.8%	4.5%	5.9%	4.1%	4.1%	2.7%	8.2%	0.5%	4.5%	1.4%	1.4%	1.8%	1.4%	1.4%
Emmet Stagg	21.5%	28.0%	6.0%	15.1%	10.2%	2.4%	2.1%	2.4%	1.5%	2.0%	1.5%	1.8%	1.0%	1.5%	0.5%	0.7%	0.8%	0.8%
Eric J. Byrne	26.7%	22.9%	7.1%	2.9%	4.3%	3.3%	2.4%	3.8%	1.0%	1.4%	1.9%	1.0%	1.9%	6.7%	1.4%	4.8%	6.2%	0.5%
Gerald Nash	35.5%	10.8%	14.4%	7.6%	5.4%	3.5%	3.5%	1.6%	1.1%	1.1%	3.0%	0.5%	1.6%	3.5%	1.6%	2.4%	2.7%	
Jack Wall	30.9%	24.2%	11.1%	4.1%	4.9%	1.9%	3.1%	3.1%	2.2%	5.8%	1.6%	2.4%	0.9%	0.7%	0.8%	1.4%		0.7%
Jan O'Sullivan	21.8%	14.0%	14.7%	8.1%	9.0%	2.7%	6.1%	1.8%	2.0%	4.8%	4.0%	2.4%	0.5%	1.1%	0.7%	4.9%	0.7%	0.8%
Joan Burton	19.4%	13.9%	14.7%	8.9%	5.8%	9.7%	2.8%	2.6%	2.7%	2.3%	3.0%	2.1%	1.5%	1.9%	1.9%	1.8%	2.7%	2.1%
Joanna Tuffy	25.1%	22.8%	6.8%	8.2%	10.1%	2.1%	2.4%	3.1%	3.0%	3.2%	3.9%	1.1%	0.8%	2.2%	1.3%	1.6%	1.3%	0.8%
Joe Costello	21.8%	15.6%	8.2%	7.8%	15.0%	4.1%	2.2%	4.6%	2.6%	2.5%	2.0%	2.5%	5.7%	1.8%	1.4%	0.8%	0.7%	0.8%
Joe Sherlock	38.0%	4.7%	14.0%	12.7%	9.3%		7.3%	4.7%	1.3%	0.7%		0.7%	5.3%			1.3%		
John Lyons	24.2%	27.8%	13.6%	4.0%	3.5%	1.0%	0.5%	2.0%	0.5%	2.5%	3.5%	1.5%	1.5%	1.5%	3.0%	4.0%	5.1%	
Kathleen Lynch	27.2%	12.8%	16.5%	6.6%	8.3%	2.3%	11.6%	1.4%	1.7%	2.4%	2.6%	1.1%	2.2%	0.4%	1.1%	0.7%	0.7%	0.4%
Kevin Humphreys	22.5%	17.5%	20.2%	9.4%	7.9%	2.5%	1.2%	2.5%	1.2%	0.2%	1.3%	0.4%	1.5%	3.5%	1.3%	3.1%	3.7%	
Liz McManus	27.1%	13.6%	12.3%	10.0%	12.8%	1.3%	5.1%	3.3%	3.5%	2.0%	1.5%	2.6%	0.6%	1.2%	1.7%	0.6%		0.8%
Mary Upton	17.6%	13.1%	11.5%	5.0%	6.3%	5.0%	2.8%	4.4%	5.7%	9.5%	2.4%	5.3%	1.4%	2.0%	4.8%	1.0%		2.4%
Michael Conaghan	13.3%	6.7%	20.0%	9.3%	6.7%	2.7%	1.3%	6.7%	1.3%	8.0%	2.7%		6.7%	4.0%	5.3%	4.0%	1.3%	
Michael D. Higgins	27.9%	16.8%	10.6%	11.2%	10.7%	5.3%	2.0%	2.6%	2.1%	2.7%	1.3%	1.9%	1.8%	0.3%	1.6%	0.4%		1.0%
Michael McCarthy	15.3%	31.3%	3.0%	6.5%	13.5%	4.8%	1.8%	5.0%	1.5%	2.0%	3.0%	0.3%	0.3%	8.3%	0.8%	0.3%	1.3%	1.5%
Michael McNama..	24.4%	17.0%	6.4%	10.0%	10.8%	3.9%	3.6%	2.3%	1.5%	1.5%	2.8%	0.5%	2.1%	3.1%	0.8%	2.3%	6.7%	0.3%
Pat Rabbitte	26.0%	14.3%	7.1%	12.6%	10.0%	4.0%	2.7%	2.9%	3.3%	1.9%	1.1%	2.8%	3.5%	2.9%	2.1%	1.3%	0.8%	0.7%
Patrick Nulty	18.0%	8.8%	10.3%	10.8%	12.9%	4.6%	3.6%	5.7%	2.1%		6.7%		2.1%	7.2%	2.6%	1.5%	2.1%	1.0%
Robert Dowds	19.9%	13.4%	11.3%	7.5%	9.1%	3.5%	3.8%	4.8%	1.6%	1.3%	1.9%	0.3%	1.9%	7.0%	3.8%	3.0%	5.9%	
Róisín Shortall	26.5%	13.9%	12.3%	12.3%	6.9%	2.5%	4.1%	2.0%	1.9%	2.7%	2.3%	3.3%	2.0%	1.1%	0.8%	1.5%	1.5%	2.2%
Ruairí Quinn	21.2%	14.3%	10.6%	8.1%	10.4%	3.9%	2.1%	2.6%	2.3%	10.7%	5.0%	2.7%	1.3%	1.5%	1.2%	0.6%	0.4%	1.1%
Séamus Pattison	46.8%	15.3%	5.0%	7.2%	11.9%		4.0%	0.1%	2.2%	2.7%		0.6%	2.8%	1.4%				
Seán Kenny	28.3%	33.8%	7.0%	3.5%	4.5%	4.3%	1.8%	1.8%	2.5%	0.8%	3.0%	1.0%	2.3%	1.0%	0.8%	2.0%	2.0%	
Seán Ryan	24.8%	5.3%	23.0%	26.5%	8.0%		4.4%	0.9%	4.4%	1.8%			0.9%					
Sean Sherlock	36.1%	13.4%	7.7%	5.2%	6.1%	3.1%	2.3%	4.3%	2.9%	5.0%	2.7%	1.0%	1.0%	1.5%	5.2%	1.3%	0.2%	1.1%
Thomas P. Broug..	26.4%	17.8%	6.6%	10.3%	7.3%	3.7%	2.1%	4.5%	3.0%	2.3%	2.2%	2.7%	2.5%	1.4%	1.5%	2.7%	1.7%	1.2%
Willie Penrose	26.6%	15.0%	15.2%	7.1%	7.6%	4.8%	2.6%	3.9%	2.2%	5.9%	0.9%	2.0%	0.8%	2.0%	0.9%	0.7%	0.7%	1.3%

% Total speeches made..

0.1% 46.8%

Figure A.1: Labour -TD by topic coverage % of total speeches made by a TD

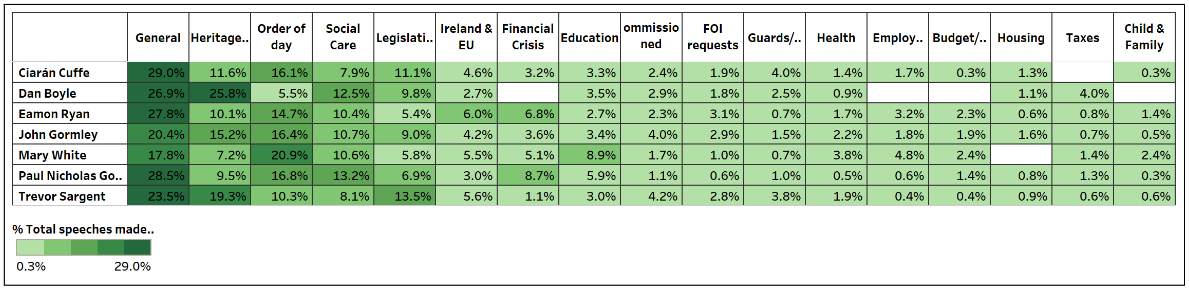


	General	Order of day	Heritage..	Social Care	Legislati..	Financial Crisis	Ireland & EU	Child & Family	Health	Taxes	ommissio ned	Water	Guards/..	Housing	Employ..	Education	FOI requests	Budget/..
Aengus Ó Snodai..	30.1%	14.5%	9.1%	12.2%	9.3%	1.7%	2.3%	2.2%	2.0%	2.2%	1.7%	1.6%	3.0%	1.8%	1.9%	1.9%	1.1%	1.4%
Arthur Morgan	30.1%	15.8%	10.5%	10.6%	5.5%	8.5%	3.4%	1.5%	2.5%	1.1%	1.5%		0.8%	1.1%	1.8%	1.4%	2.1%	1.8%
Brian Stanley	22.5%	13.0%	13.4%	8.2%	6.9%	2.8%	2.8%	1.6%	2.0%	5.0%	1.6%	10.7%	0.9%	5.1%	1.7%	0.9%	0.5%	0.6%
Caoimhghín Ó Ca..	22.8%	18.8%	9.4%	7.6%	8.7%	2.8%	2.8%	5.5%	6.9%	0.9%	3.8%	0.4%	1.3%	0.6%	1.4%	1.6%	3.3%	1.3%
Dessie Ellis	22.4%	10.1%	9.3%	12.9%	9.3%	4.0%	2.3%	3.0%	3.5%	3.2%	1.7%	2.9%	1.0%	10.3%	1.6%	1.0%	0.3%	1.3%
Gerry Adams	17.6%	14.1%	18.1%	6.5%	8.1%	5.6%	6.0%	3.0%	4.6%	4.1%	2.0%	3.1%	2.1%	1.0%	1.5%	0.9%	1.0%	0.8%
Jonathan O'Brien	22.9%	14.6%	7.8%	9.9%	12.1%	3.0%	1.7%	3.8%	2.4%	3.5%	2.4%	2.3%	1.2%	3.3%	1.3%	6.1%	0.7%	1.1%
Martin Ferris	25.6%	17.9%	9.3%	14.2%	6.2%	2.5%	3.3%	2.4%	1.6%	2.0%	1.4%	1.9%	1.4%	1.6%	3.1%	3.2%	1.6%	0.9%
Mary Lou McDon..	20.8%	13.2%	14.2%	8.4%	8.3%	3.3%	3.4%	4.7%	2.8%	4.8%	2.5%	3.2%	3.2%	1.7%	1.9%	1.0%	1.1%	1.3%
Michael Colreavy	21.2%	14.5%	11.9%	13.9%	6.8%	1.8%	4.3%	2.8%	3.1%	2.8%	2.2%	3.9%	1.8%	1.9%	3.9%	1.9%	0.7%	0.4%
Pádraig Mac Loch..	26.5%	17.1%	8.7%	8.2%	6.9%	2.5%	5.6%	3.0%	1.8%	3.5%	2.9%	1.1%	8.2%	1.0%	0.7%	1.1%	1.0%	0.3%
Peadar Tóibín	25.2%	11.5%	16.2%	7.3%	6.0%	4.8%	6.0%	2.7%	1.5%	2.6%	0.6%	3.1%	0.5%	2.3%	7.6%	1.5%	0.6%	0.1%
Pearse Doherty	19.1%	12.1%	12.5%	8.2%	10.3%	12.0%	4.2%	1.5%	1.5%	6.4%	1.3%	2.1%	1.5%	0.9%	2.5%	0.9%	1.0%	1.7%
Sandra McLellan	17.9%	12.8%	12.2%	12.3%	4.7%	4.1%	3.9%	7.7%	3.6%	2.3%	1.4%	1.7%	1.4%	5.9%	3.2%	2.6%	1.2%	1.4%
Seán Crowe	20.3%	10.3%	10.3%	18.3%	7.4%	2.7%	4.5%	5.0%	1.8%	1.1%	2.4%	0.4%	2.9%	4.0%	1.4%	5.7%	1.0%	0.3%

% Total speeches made..

0.1% 30.1%

Figure A.2: Sinn Féin -TD by topic coverage % of total speeches made by a TD

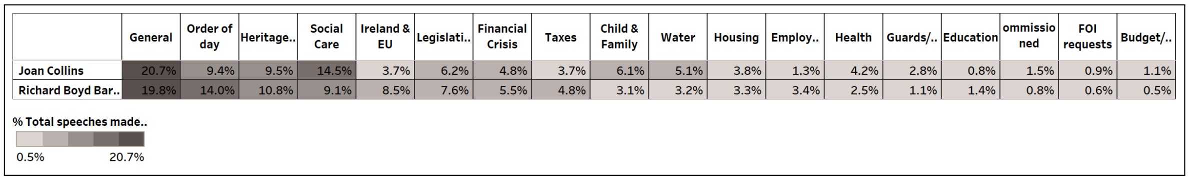


	General	Heritage..	Order of day	Social Care	Legislati..	Ireland & EU	Financial Crisis	Education	ommissio ned	FOI requests	Guards/..	Health	Employ..	Budget/..	Housing	Taxes	Child & Family
Ciarán Cuffe	29.0%	11.6%	16.1%	7.9%	11.1%	4.6%	3.2%	3.3%	2.4%	1.9%	4.0%	1.4%	1.7%	0.3%	1.3%		0.3%
Dan Boyle	26.9%	25.8%	5.5%	12.5%	9.8%	2.7%		3.5%	2.9%	1.8%	2.5%	0.9%			1.1%	4.0%	
Eamon Ryan	27.8%	10.1%	14.7%	10.4%	5.4%	6.0%	6.8%	2.7%	2.3%	3.1%	0.7%	1.7%	3.2%	2.3%	0.6%	0.8%	1.4%
John Gormley	20.4%	15.2%	16.4%	10.7%	9.0%	4.2%	3.6%	3.4%	4.0%	2.9%	1.5%	2.2%	1.8%	1.9%	1.6%	0.7%	0.5%
Mary White	17.8%	7.2%	20.9%	10.6%	5.8%	5.5%	5.1%	8.9%	1.7%	1.0%	0.7%	3.8%	4.8%	2.4%		1.4%	2.4%
Paul Nicholas Go..	28.5%	9.5%	16.8%	13.2%	6.9%	3.0%	8.7%	5.9%	1.1%	0.6%	1.0%	0.5%	0.6%	1.4%	0.8%	1.3%	0.3%
Trevor Sargent	23.5%	19.3%	10.3%	8.1%	13.5%	5.6%	1.1%	3.0%	4.2%	2.8%	3.8%	1.9%	0.4%	0.4%	0.9%	0.6%	0.6%

% Total speeches made..

0.3% 29.0%

Figure A.3: Green Party -TD by topic coverage % of total speeches made by a TD

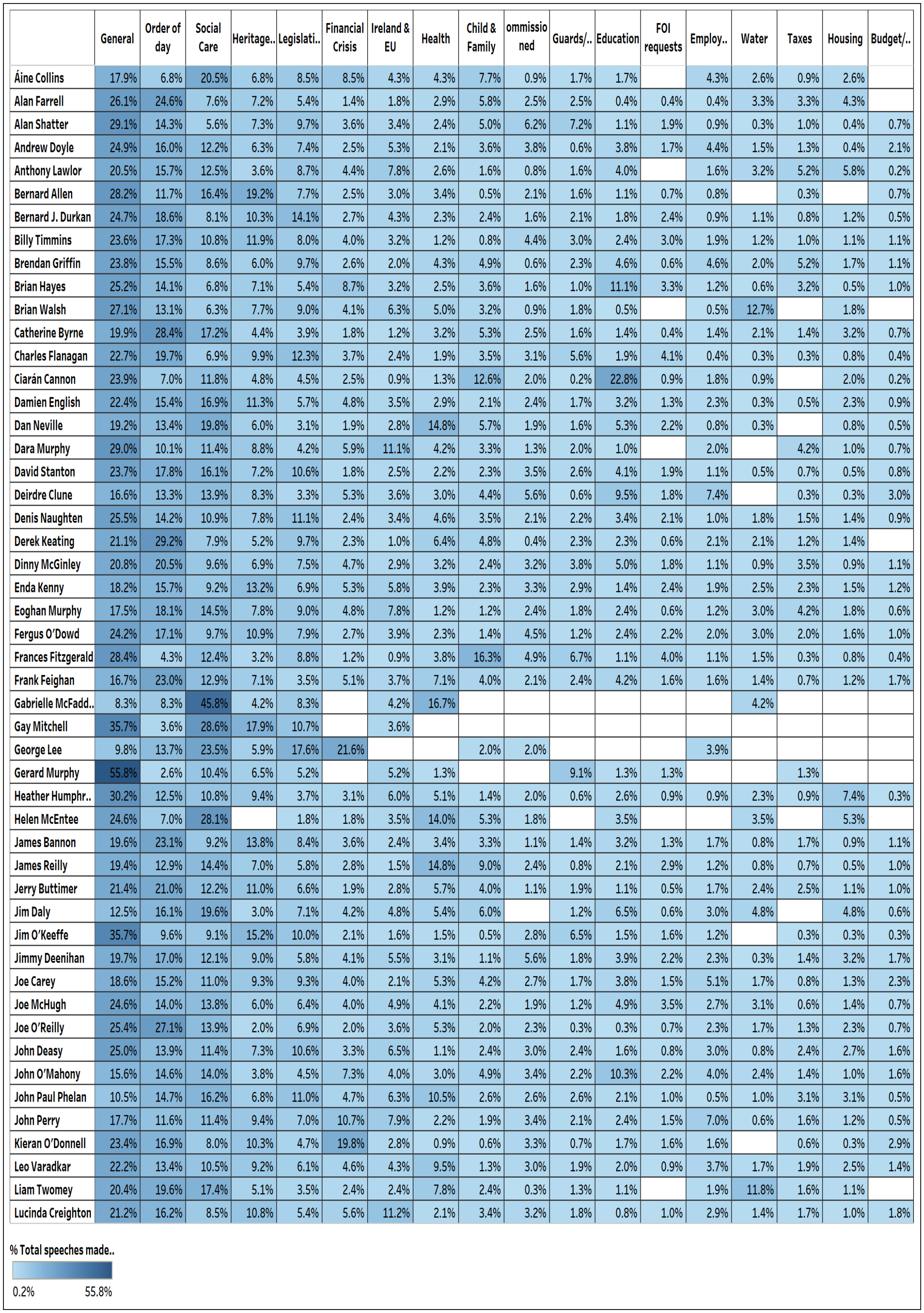


	General	Order of day	Heritage..	Social Care	Ireland & EU	Legislati..	Financial Crisis	Taxes	Child & Family	Water	Housing	Employ..	Health	Guards/..	Education	ommissio ned	FOI requests	Budget/..
Joan Collins	20.7%	9.4%	9.5%	14.5%	3.7%	6.2%	4.8%	3.7%	6.1%	5.1%	3.8%	1.3%	4.2%	2.8%	0.8%	1.5%	0.9%	1.1%
Richard Boyd Bar..	19.8%	14.0%	10.8%	9.1%	8.5%	7.6%	5.5%	4.8%	3.1%	3.2%	3.3%	3.4%	2.5%	1.1%	1.4%	0.8%	0.6%	0.5%

% Total speeches made..

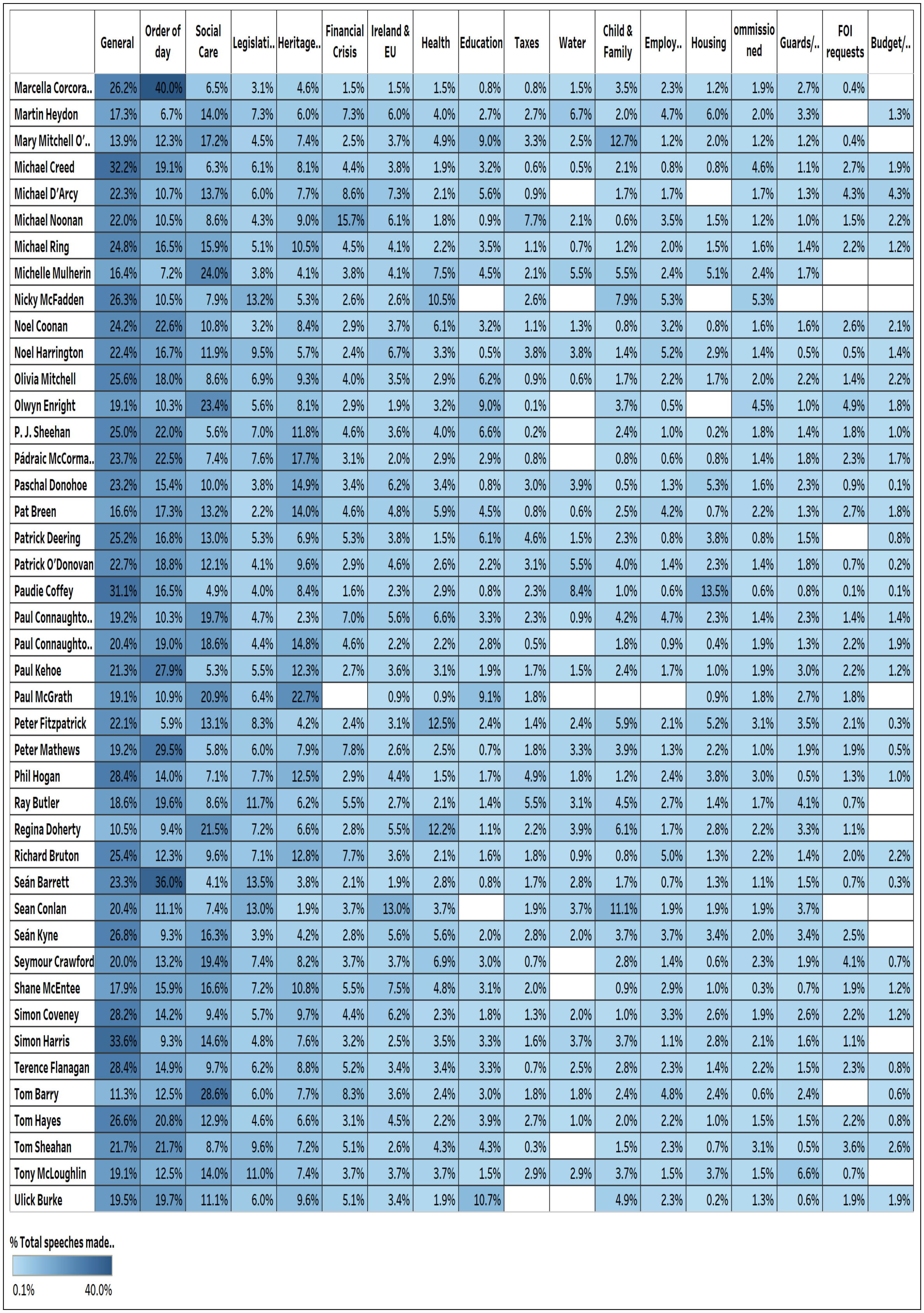
0.5% 20.7%

Figure A.4: People Before Profit -TD by topic coverage % of total speeches made by a TD



	General	Order of day	Social Care	Heritage..	Legislati..	Financial Crisis	Ireland & EU	Health	Child & Family	ommissio ned	Guards/..	Education	FOI requests	Employ..	Water	Taxes	Housing	Budget/..
Áine Collins	17.9%	6.8%	20.5%	6.8%	8.5%	8.5%	4.3%	4.3%	7.7%	0.9%	1.7%	1.7%		4.3%	2.6%	0.9%	2.6%	
Alan Farrell	26.1%	24.6%	7.6%	7.2%	5.4%	1.4%	1.8%	2.9%	5.8%	2.5%	2.5%	0.4%	0.4%	0.4%	3.3%	3.3%	4.3%	
Alan Shatter	29.1%	14.3%	5.6%	7.3%	9.7%	3.6%	3.4%	2.4%	5.0%	6.2%	7.2%	1.1%	1.9%	0.9%	0.3%	1.0%	0.4%	0.7%
Andrew Doyle	24.9%	16.0%	12.2%	6.3%	7.4%	2.5%	5.3%	2.1%	3.6%	3.8%	0.6%	3.8%	1.7%	4.4%	1.5%	1.3%	0.4%	2.1%
Anthony Lawlor	20.5%	15.7%	12.5%	3.6%	8.7%	4.4%	7.8%	2.6%	1.6%	0.8%	1.6%	4.0%		1.6%	3.2%	5.2%	5.8%	0.2%
Bernard Allen	28.2%	11.7%	16.4%	19.2%	7.7%	2.5%	3.0%	3.4%	0.5%	2.1%	1.6%	1.1%	0.7%	0.8%		0.3%		0.7%
Bernard J. Durkan	24.7%	18.6%	8.1%	10.3%	14.1%	2.7%	4.3%	2.3%	2.4%	1.6%	2.1%	1.8%	2.4%	0.9%	1.1%	0.8%	1.2%	0.5%
Billy Timmins	23.6%	17.3%	10.8%	11.9%	8.0%	4.0%	3.2%	1.2%	0.8%	4.4%	3.0%	2.4%	3.0%	1.9%	1.2%	1.0%	1.1%	1.1%
Brendan Griffin	23.8%	15.5%	8.6%	6.0%	9.7%	2.6%	2.0%	4.3%	4.9%	0.6%	2.3%	4.6%	0.6%	4.6%	2.0%	5.2%	1.7%	1.1%
Brian Hayes	25.2%	14.1%	6.8%	7.1%	5.4%	8.7%	3.2%	2.5%	3.6%	1.6%	1.0%	11.1%	3.3%	1.2%	0.6%	3.2%	0.5%	1.0%
Brian Walsh	27.1%	13.1%	6.3%	7.7%	9.0%	4.1%	6.3%	5.0%	3.2%	0.9%	1.8%	0.5%		0.5%	12.7%		1.8%	
Catherine Byrne	19.9%	28.4%	17.2%	4.4%	3.9%	1.8%	1.2%	3.2%	5.3%	2.5%	1.6%	1.4%	0.4%	1.4%	2.1%	1.4%	3.2%	0.7%
Charles Flanagan	22.7%	19.7%	6.9%	9.9%	12.3%	3.7%	2.4%	1.9%	3.5%	3.1%	5.6%	1.9%	4.1%	0.4%	0.3%	0.3%	0.8%	0.4%
Ciarán Cannon	23.9%	7.0%	11.8%	4.8%	4.5%	2.5%	0.9%	1.3%	12.6%	2.0%	0.2%	22.8%	0.9%	1.8%	0.9%		2.0%	0.2%
Damien English	22.4%	15.4%	16.9%	11.3%	5.7%	4.8%	3.5%	2.9%	2.1%	2.4%	1.7%	3.2%	1.3%	2.3%	0.3%	0.5%	2.3%	0.9%
Dan Neville	19.2%	13.4%	19.8%	6.0%	3.1%	1.9%	2.8%	14.8%	5.7%	1.9%	1.6%	5.3%	2.2%	0.8%	0.3%		0.8%	0.5%
Dara Murphy	29.0%	10.1%	11.4%	8.8%	4.2%	5.9%	11.1%	4.2%	3.3%	1.3%	2.0%	1.0%		2.0%		4.2%	1.0%	0.7%
David Stanton	23.7%	17.8%	16.1%	7.2%	10.6%	1.8%	2.5%	2.2%	2.3%	3.5%	2.6%	4.1%	1.9%	1.1%	0.5%	0.7%	0.5%	0.8%
Deirdre Clune	16.6%	13.3%	13.9%	8.3%	3.3%	5.3%	3.6%	3.0%	4.4%	5.6%	0.6%	9.5%	1.8%	7.4%		0.3%	0.3%	3.0%
Denis Naughten	25.5%	14.2%	10.9%	7.8%	11.1%	2.4%	3.4%	4.6%	3.5%	2.1%	2.2%	3.4%	2.1%	1.0%	1.8%	1.5%	1.4%	0.9%
Derek Keating	21.1%	29.2%	7.9%	5.2%	9.7%	2.3%	1.0%	6.4%	4.8%	0.4%	2.3%	2.3%	0.6%	2.1%	2.1%	1.2%	1.4%	
Dinny McGinley	20.8%	20.5%	9.6%	6.9%	7.5%	4.7%	2.9%	3.2%	2.4%	3.2%	3.8%	5.0%	1.8%	1.1%	0.9%	3.5%	0.9%	1.1%
Enda Kenny	18.2%	15.7%	9.2%	13.2%	6.9%	5.3%	5.8%	3.9%	2.3%	3.3%	2.9%	1.4%	2.4%	1.9%	2.5%	2.3%	1.5%	1.2%
Eoghan Murphy	17.5%	18.1%	14.5%	7.8%	9.0%	4.8%	7.8%	1.2%	1.2%	2.4%	1.8%	2.4%	0.6%	1.2%	3.0%	4.2%	1.8%	0.6%
Fergus O'Dowd	24.2%	17.1%	9.7%	10.9%	7.9%	2.7%	3.9%	2.3%	1.4%	4.5%	1.2%	2.4%	2.2%	2.0%	3.0%	2.0%	1.6%	1.0%
Frances Fitzgerald	28.4%	4.3%	12.4%	3.2%	8.8%	1.2%	0.9%	3.8%	16.3%	4.9%	6.7%	1.1%	4.0%	1.1%	1.5%	0.3%	0.8%	0.4%
Frank Feighan	16.7%	23.0%	12.9%	7.1%	3.5%	5.1%	3.7%	7.1%	4.0%	2.1%	2.4%	4.2%	1.6%	1.6%	1.4%	0.7%	1.2%	1.7%
Gabrielle McFadd..	8.3%	8.3%	45.8%	4.2%	8.3%		4.2%	16.7%							4.2%			
Gay Mitchell	35.7%	3.6%	28.6%	17.9%	10.7%		3.6%											
George Lee	9.8%	13.7%	23.5%	5.9%	17.6%	21.6%			2.0%	2.0%				3.9%				
Gerard Murphy	55.8%	2.6%	10.4%	6.5%	5.2%		5.2%	1.3%			9.1%	1.3%	1.3%			1.3%		
Heather Humphr..	30.2%	12.5%	10.8%	9.4%	3.7%	3.1%	6.0%	5.1%	1.4%	2.0%	0.6%	2.6%	0.9%	0.9%	2.3%	0.9%	7.4%	0.3%
Helen McEntee	24.6%	7.0%	28.1%		1.8%	1.8%	3.5%	14.0%	5.3%	1.8%		3.5%			3.5%		5.3%	
James Bannon	19.6%	23.1%	9.2%	13.8%	8.4%	3.6%	2.4%	3.4%	3.3%	1.1%	1.4%	3.2%	1.3%	1.7%	0.8%	1.7%	0.9%	1.1%
James Reilly	19.4%	12.9%	14.4%	7.0%	5.8%	2.8%	1.5%	14.8%	9.0%	2.4%	0.8%	2.1%	2.9%	1.2%	0.8%	0.7%	0.5%	1.0%
Jerry Buttimer	21.4%	21.0%	12.2%	11.0%	6.6%	1.9%	2.8%	5.7%	4.0%	1.1%	1.9%	1.1%	0.5%	1.7%	2.4%	2.5%	1.1%	1.0%
Jim Daly	12.5%	16.1%	19.6%	3.0%	7.1%	4.2%	4.8%	5.4%	6.0%		1.2%	6.5%	0.6%	3.0%	4.8%		4.8%	0.6%
Jim O'Keeffe	35.7%	9.6%	9.1%	15.2%	10.0%	2.1%	1.6%	1.5%	0.5%	2.8%	6.5%	1.5%	1.6%	1.2%		0.3%	0.3%	0.3%
Jimmy Deenihan	19.7%	17.0%	12.1%	9.0%	5.8%	4.1%	5.5%	3.1%	1.1%	5.6%	1.8%	3.9%	2.2%	2.3%	0.3%	1.4%	3.2%	1.7%
Joe Carey	18.6%	15.2%	11.0%	9.3%	9.3%	4.0%	2.1%	5.3%	4.2%	2.7%	1.7%	3.8%	1.5%	5.1%	1.7%	0.8%	1.3%	2.3%
Joe McHugh	24.6%	14.0%	13.8%	6.0%	6.4%	4.0%	4.9%	4.1%	2.2%	1.9%	1.2%	4.9%	3.5%	2.7%	3.1%	0.6%	1.4%	0.7%
Joe O'Reilly	25.4%	27.1%	13.9%	2.0%	6.9%	2.0%	3.6%	5.3%	2.0%	2.3%	0.3%	0.3%	0.7%	2.3%	1.7%	1.3%	2.3%	0.7%
John Deasy	25.0%	13.9%	11.4%	7.3%	10.6%	3.3%	6.5%	1.1%	2.4%	3.0%	2.4%	1.6%	0.8%	3.0%	0.8%	2.4%	2.7%	1.6%
John O'Mahony	15.6%	14.6%	14.0%	3.8%	4.5%	7.3%	4.0%	3.0%	4.9%	3.4%	2.2%	10.3%	2.2%	4.0%	2.4%	1.4%	1.0%	1.6%
John Paul Phelan	10.5%	14.7%	16.2%	6.8%	11.0%	4.7%	6.3%	10.5%	2.6%	2.6%	2.6%	2.1%	1.0%	0.5%	1.0%	3.1%	3.1%	0.5%
John Perry	17.7%	11.6%	11.4%	9.4%	7.0%	10.7%	7.9%	2.2%	1.9%	3.4%	2.1%	2.4%	1.5%	7.0%	0.6%	1.6%	1.2%	0.5%
Kieran O'Donnell	23.4%	16.9%	8.0%	10.3%	4.7%	19.8%	2.8%	0.9%	0.6%	3.3%	0.7%	1.7%	1.6%	1.6%		0.6%	0.3%	2.9%
Leo Varadkar	22.2%	13.4%	10.5%	9.2%	6.1%	4.6%	4.3%	9.5%	1.3%	3.0%	1.9%	2.0%	0.9%	3.7%	1.7%	1.9%	2.5%	1.4%
Liam Twomey	20.4%	19.6%	17.4%	5.1%	3.5%	2.4%	2.4%	7.8%	2.4%	0.3%	1.3%	1.1%		1.9%	11.8%	1.6%	1.1%	
Lucinda Creighton	21.2%	16.2%	8.5%	10.8%	5.4%	5.6%	11.2%	2.1%	3.4%	3.2%	1.8%	0.8%	1.0%	2.9%	1.4%	1.7%	1.0%	1.8%

Figure A.5: Fine Gael -TD by topic coverage % of total speeches made by a TD

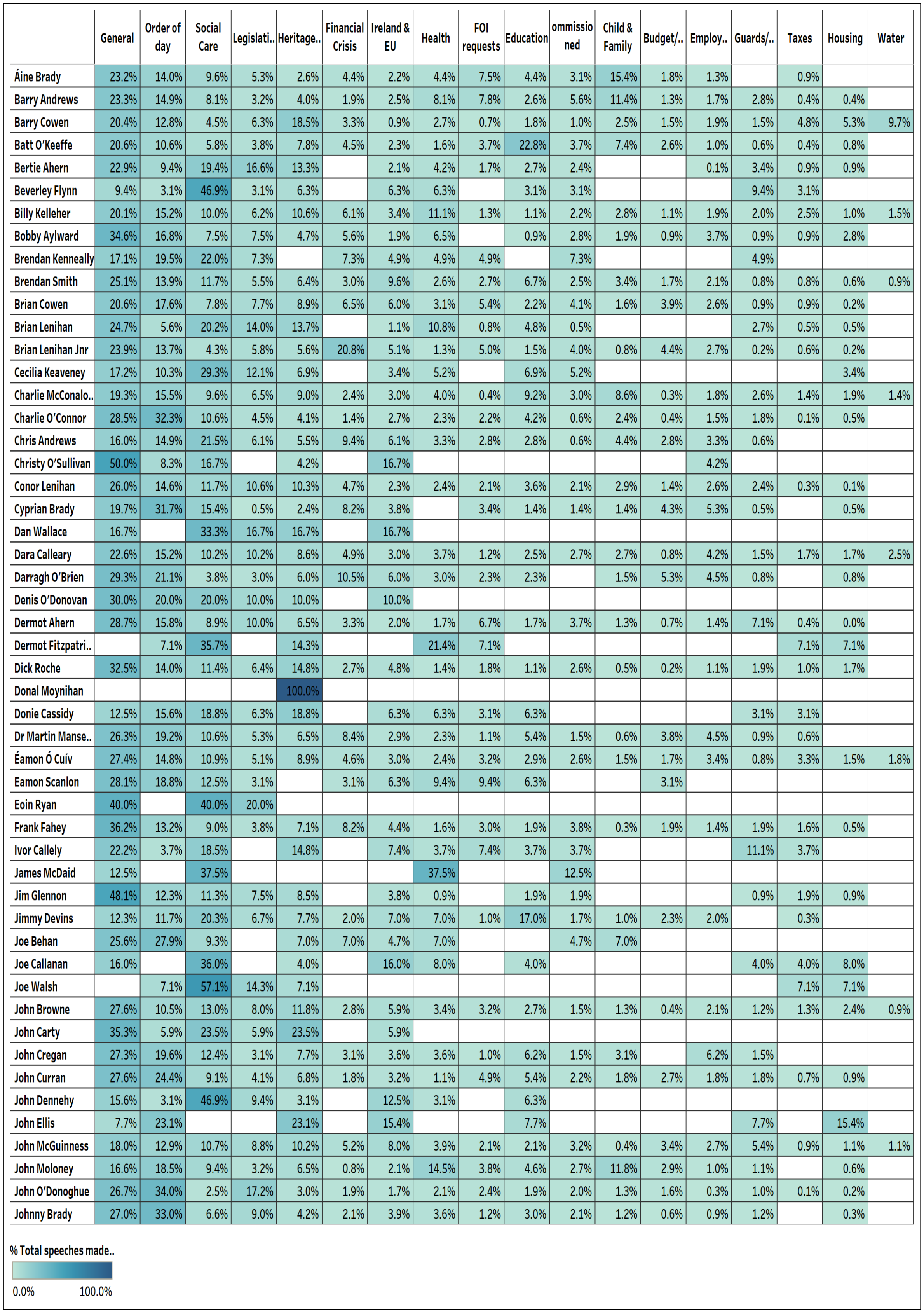


	General	Order of day	Social Care	Legislati..	Heritage..	Financial Crisis	Ireland & EU	Health	Education	Taxes	Water	Child & Family	Employ..	Housing	ommissio ned	Guards/..	FOI requests	Budget/..
Marcella Corcora..	26.2%	40.0%	6.5%	3.1%	4.6%	1.5%	1.5%	1.5%	0.8%	0.8%	1.5%	3.5%	2.3%	1.2%	1.9%	2.7%	0.4%	
Martin Heydon	17.3%	6.7%	14.0%	7.3%	6.0%	7.3%	6.0%	4.0%	2.7%	2.7%	6.7%	2.0%	4.7%	6.0%	2.0%	3.3%		1.3%
Mary Mitchell O'..	13.9%	12.3%	17.2%	4.5%	7.4%	2.5%	3.7%	4.9%	9.0%	3.3%	2.5%	12.7%	1.2%	2.0%	1.2%	1.2%	0.4%	
Michael Creed	32.2%	19.1%	6.3%	6.1%	8.1%	4.4%	3.8%	1.9%	3.2%	0.6%	0.5%	2.1%	0.8%	0.8%	4.6%	1.1%	2.7%	1.9%
Michael D'Arcy	22.3%	10.7%	13.7%	6.0%	7.7%	8.6%	7.3%	2.1%	5.6%	0.9%		1.7%	1.7%		1.7%	1.3%	4.3%	4.3%
Michael Noonan	22.0%	10.5%	8.6%	4.3%	9.0%	15.7%	6.1%	1.8%	0.9%	7.7%	2.1%	0.6%	3.5%	1.5%	1.2%	1.0%	1.5%	2.2%
Michael Ring	24.8%	16.5%	15.9%	5.1%	10.5%	4.5%	4.1%	2.2%	3.5%	1.1%	0.7%	1.2%	2.0%	1.5%	1.6%	1.4%	2.2%	1.2%
Michelle Mulherin	16.4%	7.2%	24.0%	3.8%	4.1%	3.8%	4.1%	7.5%	4.5%	2.1%	5.5%	5.5%	2.4%	5.1%	2.4%	1.7%		
Nicky McFadden	26.3%	10.5%	7.9%	13.2%	5.3%	2.6%	2.6%	10.5%		2.6%		7.9%	5.3%		5.3%			
Noel Coonan	24.2%	22.6%	10.8%	3.2%	8.4%	2.9%	3.7%	6.1%	3.2%	1.1%	1.3%	0.8%	3.2%	0.8%	1.6%	1.6%	2.6%	2.1%
Noel Harrington	22.4%	16.7%	11.9%	9.5%	5.7%	2.4%	6.7%	3.3%	0.5%	3.8%	3.8%	1.4%	5.2%	2.9%	1.4%	0.5%	0.5%	1.4%
Olivia Mitchell	25.6%	18.0%	8.6%	6.9%	9.3%	4.0%	3.5%	2.9%	6.2%	0.9%	0.6%	1.7%	2.2%	1.7%	2.0%	2.2%	1.4%	2.2%
Olwyn Enright	19.1%	10.3%	23.4%	5.6%	8.1%	2.9%	1.9%	3.2%	9.0%	0.1%		3.7%	0.5%		4.5%	1.0%	4.9%	1.8%
P. J. Sheehan	25.0%	22.0%	5.6%	7.0%	11.8%	4.6%	3.6%	4.0%	6.6%	0.2%		2.4%	1.0%	0.2%	1.8%	1.4%	1.8%	1.0%
Pádraic McCorma..	23.7%	22.5%	7.4%	7.6%	17.7%	3.1%	2.0%	2.9%	2.9%	0.8%		0.8%	0.6%	0.8%	1.4%	1.8%	2.3%	1.7%
Paschal Donohoe	23.2%	15.4%	10.0%	3.8%	14.9%	3.4%	6.2%	3.4%	0.8%	3.0%	3.9%	0.5%	1.3%	5.3%	1.6%	2.3%	0.9%	0.1%
Pat Breen	16.6%	17.3%	13.2%	2.2%	14.0%	4.6%	4.8%	5.9%	4.5%	0.8%	0.6%	2.5%	4.2%	0.7%	2.2%	1.3%	2.7%	1.8%
Patrick Deering	25.2%	16.8%	13.0%	5.3%	6.9%	5.3%	3.8%	1.5%	6.1%	4.6%	1.5%	2.3%	0.8%	3.8%	0.8%	1.5%		0.8%
Patrick O'Donovan	22.7%	18.8%	12.1%	4.1%	9.6%	2.9%	4.6%	2.6%	2.2%	3.1%	5.5%	4.0%	1.4%	2.3%	1.4%	1.8%	0.7%	0.2%
Paudie Coffey	31.1%	16.5%	4.9%	4.0%	8.4%	1.6%	2.3%	2.9%	0.8%	2.3%	8.4%	1.0%	0.6%	13.5%	0.6%	0.8%	0.1%	0.1%
Paul Connaughto..	19.2%	10.3%	19.7%	4.7%	2.3%	7.0%	5.6%	6.6%	3.3%	2.3%	0.9%	4.2%	4.7%	2.3%	1.4%	2.3%	1.4%	1.4%
Paul Connaughto..	20.4%	19.0%	18.6%	4.4%	14.8%	4.6%	2.2%	2.2%	2.8%	0.5%		1.8%	0.9%	0.4%	1.9%	1.3%	2.2%	1.9%
Paul Kehoe	21.3%	27.9%	5.3%	5.5%	12.3%	2.7%	3.6%	3.1%	1.9%	1.7%	1.5%	2.4%	1.7%	1.0%	1.9%	3.0%	2.2%	1.2%
Paul McGrath	19.1%	10.9%	20.9%	6.4%	22.7%		0.9%	0.9%	9.1%	1.8%				0.9%	1.8%	2.7%	1.8%	
Peter Fitzpatrick	22.1%	5.9%	13.1%	8.3%	4.2%	2.4%	3.1%	12.5%	2.4%	1.4%	2.4%	5.9%	2.1%	5.2%	3.1%	3.5%	2.1%	0.3%
Peter Mathews	19.2%	29.5%	5.8%	6.0%	7.9%	7.8%	2.6%	2.5%	0.7%	1.8%	3.3%	3.9%	1.3%	2.2%	1.0%	1.9%	1.9%	0.5%
Phil Hogan	28.4%	14.0%	7.1%	7.7%	12.5%	2.9%	4.4%	1.5%	1.7%	4.9%	1.8%	1.2%	2.4%	3.8%	3.0%	0.5%	1.3%	1.0%
Ray Butler	18.6%	19.6%	8.6%	11.7%	6.2%	5.5%	2.7%	2.1%	1.4%	5.5%	3.1%	4.5%	2.7%	1.4%	1.7%	4.1%	0.7%	
Regina Doherty	10.5%	9.4%	21.5%	7.2%	6.6%	2.8%	5.5%	12.2%	1.1%	2.2%	3.9%	6.1%	1.7%	2.8%	2.2%	3.3%	1.1%	
Richard Bruton	25.4%	12.3%	9.6%	7.1%	12.8%	7.7%	3.6%	2.1%	1.6%	1.8%	0.9%	0.8%	5.0%	1.3%	2.2%	1.4%	2.0%	2.2%
Seán Barrett	23.3%	36.0%	4.1%	13.5%	3.8%	2.1%	1.9%	2.8%	0.8%	1.7%	2.8%	1.7%	0.7%	1.3%	1.1%	1.5%	0.7%	0.3%
Sean Conlan	20.4%	11.1%	7.4%	13.0%	1.9%	3.7%	13.0%	3.7%		1.9%	3.7%	11.1%	1.9%	1.9%	1.9%	3.7%		
Seán Kyne	26.8%	9.3%	16.3%	3.9%	4.2%	2.8%	5.6%	5.6%	2.0%	2.8%	2.0%	3.7%	3.7%	3.4%	2.0%	3.4%	2.5%	
Seymour Crawford	20.0%	13.2%	19.4%	7.4%	8.2%	3.7%	3.7%	6.9%	3.0%	0.7%		2.8%	1.4%	0.6%	2.3%	1.9%	4.1%	0.7%
Shane McEntee	17.9%	15.9%	16.6%	7.2%	10.8%	5.5%	7.5%	4.8%	3.1%	2.0%		0.9%	2.9%	1.0%	0.3%	0.7%	1.9%	1.2%
Simon Coveney	28.2%	14.2%	9.4%	5.7%	9.7%	4.4%	6.2%	2.3%	1.8%	1.3%	2.0%	1.0%	3.3%	2.6%	1.9%	2.6%	2.2%	1.2%
Simon Harris	33.6%	9.3%	14.6%	4.8%	7.6%	3.2%	2.5%	3.5%	3.3%	1.6%	3.7%	3.7%	1.1%	2.8%	2.1%	1.6%	1.1%	
Terence Flanagan	28.4%	14.9%	9.7%	6.2%	8.8%	5.2%	3.4%	3.4%	3.3%	0.7%	2.5%	2.8%	2.3%	1.4%	2.2%	1.5%	2.3%	0.8%
Tom Barry	11.3%	12.5%	28.6%	6.0%	7.7%	8.3%	3.6%	2.4%	3.0%	1.8%	1.8%	2.4%	4.8%	2.4%	0.6%	2.4%		0.6%
Tom Hayes	26.6%	20.8%	12.9%	4.6%	6.6%	3.1%	4.5%	2.2%	3.9%	2.7%	1.0%	2.0%	2.2%	1.0%	1.5%	1.5%	2.2%	0.8%
Tom Sheahan	21.7%	21.7%	8.7%	9.6%	7.2%	5.1%	2.6%	4.3%	4.3%	0.3%		1.5%	2.3%	0.7%	3.1%	0.5%	3.6%	2.6%
Tony McLoughlin	19.1%	12.5%	14.0%	11.0%	7.4%	3.7%	3.7%	3.7%	1.5%	2.9%	2.9%	3.7%	1.5%	3.7%	1.5%	6.6%	0.7%	
Ulick Burke	19.5%	19.7%	11.1%	6.0%	9.6%	5.1%	3.4%	1.9%	10.7%			4.9%	2.3%	0.2%	1.3%	0.6%	1.9%	1.9%

% Total speeches made..

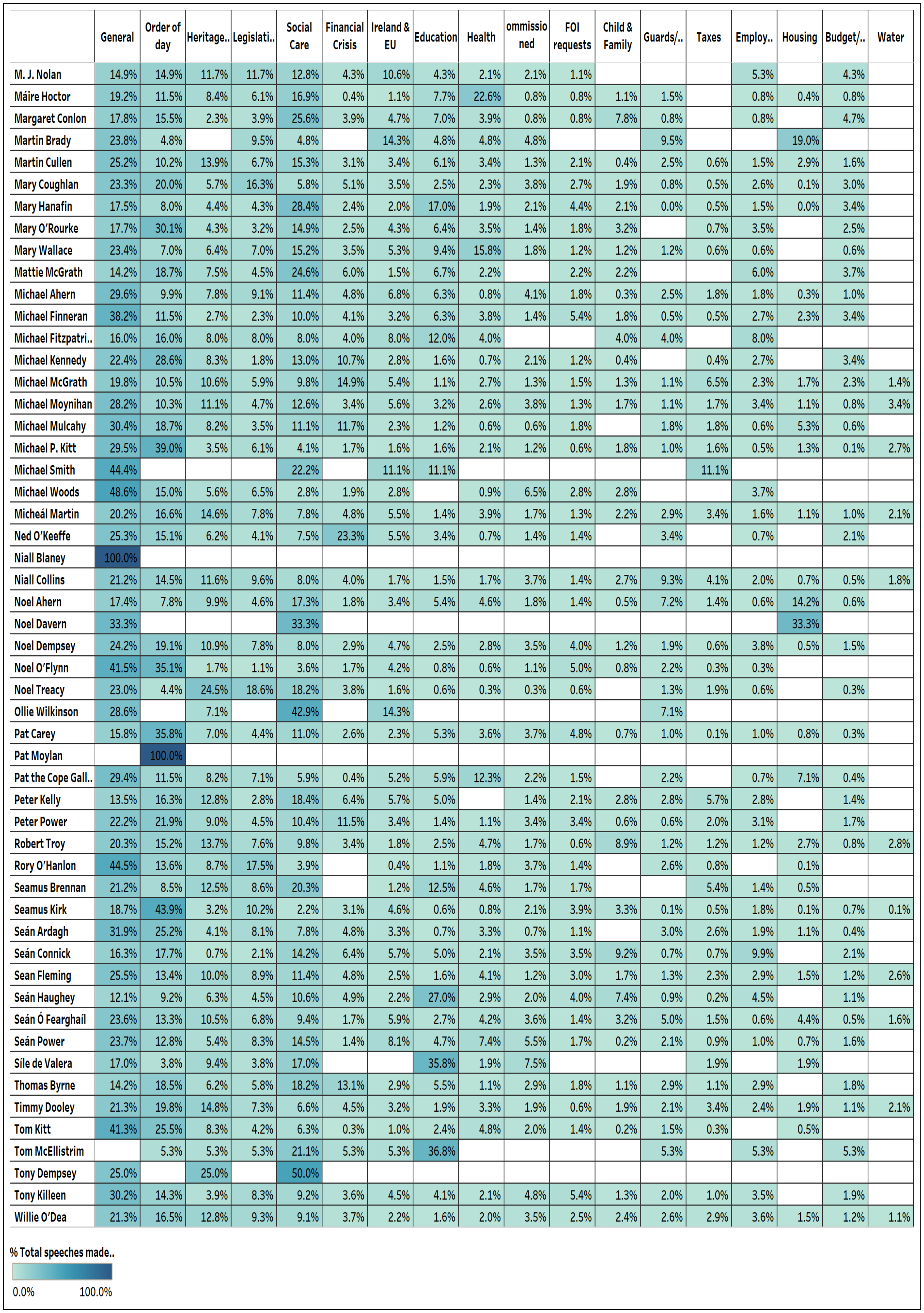
0.1% 40.0%

Figure A.6: Fine Gael - TD by topic coverage % of total speeches made by a TD (continued)



	General	Order of day	Social Care	Legislati..	Heritage..	Financial Crisis	Ireland & EU	Health	FOI requests	Education	ommissioned	Child & Family	Budget/..	Employ..	Guards/..	Taxes	Housing	Water
Áine Brady	23.2%	14.0%	9.6%	5.3%	2.6%	4.4%	2.2%	4.4%	7.5%	4.4%	3.1%	15.4%	1.8%	1.3%		0.9%		
Barry Andrews	23.3%	14.9%	8.1%	3.2%	4.0%	1.9%	2.5%	8.1%	7.8%	2.6%	5.6%	11.4%	1.3%	1.7%	2.8%	0.4%	0.4%	
Barry Cowen	20.4%	12.8%	4.5%	6.3%	18.5%	3.3%	0.9%	2.7%	0.7%	1.8%	1.0%	2.5%	1.5%	1.9%	1.5%	4.8%	5.3%	9.7%
Batt O'Keeffe	20.6%	10.6%	5.8%	3.8%	7.8%	4.5%	2.3%	1.6%	3.7%	22.8%	3.7%	7.4%	2.6%	1.0%	0.6%	0.4%	0.8%	
Bertie Ahern	22.9%	9.4%	19.4%	16.6%	13.3%		2.1%	4.2%	1.7%	2.7%	2.4%			0.1%	3.4%	0.9%	0.9%	
Beverley Flynn	9.4%	3.1%	46.9%	3.1%	6.3%		6.3%	6.3%		3.1%	3.1%				9.4%	3.1%		
Billy Kelleher	20.1%	15.2%	10.0%	6.2%	10.6%	6.1%	3.4%	11.1%	1.3%	1.1%	2.2%	2.8%	1.1%	1.9%	2.0%	2.5%	1.0%	1.5%
Bobby Aylward	34.6%	16.8%	7.5%	7.5%	4.7%	5.6%	1.9%	6.5%		0.9%	2.8%	1.9%	0.9%	3.7%	0.9%	0.9%	2.8%	
Brendan Kenneally	17.1%	19.5%	22.0%	7.3%		7.3%	4.9%	4.9%	4.9%		7.3%				4.9%			
Brendan Smith	25.1%	13.9%	11.7%	5.5%	6.4%	3.0%	9.6%	2.6%	2.7%	6.7%	2.5%	3.4%	1.7%	2.1%	0.8%	0.8%	0.6%	0.9%
Brian Cowen	20.6%	17.6%	7.8%	7.7%	8.9%	6.5%	6.0%	3.1%	5.4%	2.2%	4.1%	1.6%	3.9%	2.6%	0.9%	0.9%	0.2%	
Brian Lenihan	24.7%	5.6%	20.2%	14.0%	13.7%		1.1%	10.8%	0.8%	4.8%	0.5%				2.7%	0.5%	0.5%	
Brian Lenihan Jnr	23.9%	13.7%	4.3%	5.8%	5.6%	20.8%	5.1%	1.3%	5.0%	1.5%	4.0%	0.8%	4.4%	2.7%	0.2%	0.6%	0.2%	
Cecilia Keaveney	17.2%	10.3%	29.3%	12.1%	6.9%		3.4%	5.2%		6.9%	5.2%						3.4%	
Charlie McConalo..	19.3%	15.5%	9.6%	6.5%	9.0%	2.4%	3.0%	4.0%	0.4%	9.2%	3.0%	8.6%	0.3%	1.8%	2.6%	1.4%	1.9%	1.4%
Charlie O'Connor	28.5%	32.3%	10.6%	4.5%	4.1%	1.4%	2.7%	2.3%	2.2%	4.2%	0.6%	2.4%	0.4%	1.5%	1.8%	0.1%	0.5%	
Chris Andrews	16.0%	14.9%	21.5%	6.1%	5.5%	9.4%	6.1%	3.3%	2.8%	2.8%	0.6%	4.4%	2.8%	3.3%	0.6%			
Christy O'Sullivan	50.0%	8.3%	16.7%		4.2%		16.7%							4.2%				
Conor Lenihan	26.0%	14.6%	11.7%	10.6%	10.3%	4.7%	2.3%	2.4%	2.1%	3.6%	2.1%	2.9%	1.4%	2.6%	2.4%	0.3%	0.1%	
Cyprian Brady	19.7%	31.7%	15.4%	0.5%	2.4%	8.2%	3.8%		3.4%	1.4%	1.4%	1.4%	4.3%	5.3%	0.5%		0.5%	
Dan Wallace	16.7%		33.3%	16.7%	16.7%		16.7%											
Dara Calleary	22.6%	15.2%	10.2%	10.2%	8.6%	4.9%	3.0%	3.7%	1.2%	2.5%	2.7%	2.7%	0.8%	4.2%	1.5%	1.7%	1.7%	2.5%
Darragh O'Brien	29.3%	21.1%	3.8%	3.0%	6.0%	10.5%	6.0%	3.0%	2.3%	2.3%		1.5%	5.3%	4.5%	0.8%		0.8%	
Denis O'Donovan	30.0%	20.0%	20.0%	10.0%	10.0%		10.0%											
Dermot Ahern	28.7%	15.8%	8.9%	10.0%	6.5%	3.3%	2.0%	1.7%	6.7%	1.7%	3.7%	1.3%	0.7%	1.4%	7.1%	0.4%	0.0%	
Dermot Fitzpatri..		7.1%	35.7%		14.3%			21.4%	7.1%							7.1%	7.1%	
Dick Roche	32.5%	14.0%	11.4%	6.4%	14.8%	2.7%	4.8%	1.4%	1.8%	1.1%	2.6%	0.5%	0.2%	1.1%	1.9%	1.0%	1.7%	
Donal Moynihan					100.0%													
Donie Cassidy	12.5%	15.6%	18.8%	6.3%	18.8%		6.3%	6.3%	3.1%	6.3%					3.1%	3.1%		
Dr Martin Manse..	26.3%	19.2%	10.6%	5.3%	6.5%	8.4%	2.9%	2.3%	1.1%	5.4%	1.5%	0.6%	3.8%	4.5%	0.9%	0.6%		
Éamon Ó Cuív	27.4%	14.8%	10.9%	5.1%	8.9%	4.6%	3.0%	2.4%	3.2%	2.9%	2.6%	1.5%	1.7%	3.4%	0.8%	3.3%	1.5%	1.8%
Eamon Scanlon	28.1%	18.8%	12.5%	3.1%		3.1%	6.3%	9.4%	9.4%	6.3%			3.1%					
Eoin Ryan	40.0%		40.0%	20.0%														
Frank Fahey	36.2%	13.2%	9.0%	3.8%	7.1%	8.2%	4.4%	1.6%	3.0%	1.9%	3.8%	0.3%	1.9%	1.4%	1.9%	1.6%	0.5%	
Ivor Callely	22.2%	3.7%	18.5%		14.8%		7.4%	3.7%	7.4%	3.7%	3.7%				11.1%	3.7%		
James McDaid	12.5%		37.5%					37.5%			12.5%							
Jim Glennon	48.1%	12.3%	11.3%	7.5%	8.5%		3.8%	0.9%		1.9%	1.9%				0.9%	1.9%	0.9%	
Jimmy Devins	12.3%	11.7%	20.3%	6.7%	7.7%	2.0%	7.0%	7.0%	1.0%	17.0%	1.7%	1.0%	2.3%	2.0%		0.3%		
Joe Behan	25.6%	27.9%	9.3%		7.0%	7.0%	4.7%	7.0%			4.7%	7.0%						
Joe Callanan	16.0%		36.0%		4.0%		16.0%	8.0%		4.0%					4.0%	4.0%	8.0%	
Joe Walsh		7.1%	57.1%	14.3%	7.1%											7.1%	7.1%	
John Browne	27.6%	10.5%	13.0%	8.0%	11.8%	2.8%	5.9%	3.4%	3.2%	2.7%	1.5%	1.3%	0.4%	2.1%	1.2%	1.3%	2.4%	0.9%
John Carty	35.3%	5.9%	23.5%	5.9%	23.5%		5.9%											
John Cregan	27.3%	19.6%	12.4%	3.1%	7.7%	3.1%	3.6%	3.6%	1.0%	6.2%	1.5%	3.1%		6.2%	1.5%			
John Curran	27.6%	24.4%	9.1%	4.1%	6.8%	1.8%	3.2%	1.1%	4.9%	5.4%	2.2%	1.8%	2.7%	1.8%	1.8%	0.7%	0.9%	
John Dennehy	15.6%	3.1%	46.9%	9.4%	3.1%		12.5%	3.1%		6.3%								
John Ellis	7.7%	23.1%			23.1%		15.4%			7.7%					7.7%		15.4%	
John McGuinness	18.0%	12.9%	10.7%	8.8%	10.2%	5.2%	8.0%	3.9%	2.1%	2.1%	3.2%	0.4%	3.4%	2.7%	5.4%	0.9%	1.1%	1.1%
John Moloney	16.6%	18.5%	9.4%	3.2%	6.5%	0.8%	2.1%	14.5%	3.8%	4.6%	2.7%	11.8%	2.9%	1.0%	1.1%		0.6%	
John O'Donoghue	26.7%	34.0%	2.5%	17.2%	3.0%	1.9%	1.7%	2.1%	2.4%	1.9%	2.0%	1.3%	1.6%	0.3%	1.0%	0.1%	0.2%	
Johnny Brady	27.0%	33.0%	6.6%	9.0%	4.2%	2.1%	3.9%	3.6%	1.2%	3.0%	2.1%	1.2%	0.6%	0.9%	1.2%		0.3%	

Figure A.7: Fianna Fáil -TD by topic coverage % of total speeches made by a TD



	General	Order of day	Heritage..	Legislati..	Social Care	Financial Crisis	Ireland & EU	Education	Health	ommissioned	FOI requests	Child & Family	Guards/..	Taxes	Employ..	Housing	Budget/..	Water
M. J. Nolan	14.9%	14.9%	11.7%	11.7%	12.8%	4.3%	10.6%	4.3%	2.1%	2.1%	1.1%				5.3%		4.3%	
Máire Hoctor	19.2%	11.5%	8.4%	6.1%	16.9%	0.4%	1.1%	7.7%	22.6%	0.8%	0.8%	1.1%	1.5%		0.8%	0.4%	0.8%	
Margaret Conlon	17.8%	15.5%	2.3%	3.9%	25.6%	3.9%	4.7%	7.0%	3.9%	0.8%	0.8%	7.8%	0.8%		0.8%		4.7%	
Martin Brady	23.8%	4.8%		9.5%	4.8%		14.3%	4.8%	4.8%	4.8%			9.5%			19.0%		
Martin Cullen	25.2%	10.2%	13.9%	6.7%	15.3%	3.1%	3.4%	6.1%	3.4%	1.3%	2.1%	0.4%	2.5%	0.6%	1.5%	2.9%	1.6%	
Mary Coughlan	23.3%	20.0%	5.7%	16.3%	5.8%	5.1%	3.5%	2.5%	2.3%	3.8%	2.7%	1.9%	0.8%	0.5%	2.6%	0.1%	3.0%	
Mary Hanafin	17.5%	8.0%	4.4%	4.3%	28.4%	2.4%	2.0%	17.0%	1.9%	2.1%	4.4%	2.1%	0.0%	0.5%	1.5%	0.0%	3.4%	
Mary O'Rourke	17.7%	30.1%	4.3%	3.2%	14.9%	2.5%	4.3%	6.4%	3.5%	1.4%	1.8%	3.2%		0.7%	3.5%		2.5%	
Mary Wallace	23.4%	7.0%	6.4%	7.0%	15.2%	3.5%	5.3%	9.4%	15.8%	1.8%	1.2%	1.2%	1.2%	0.6%	0.6%		0.6%	
Mattie McGrath	14.2%	18.7%	7.5%	4.5%	24.6%	6.0%	1.5%	6.7%	2.2%		2.2%	2.2%			6.0%		3.7%	
Michael Ahern	29.6%	9.9%	7.8%	9.1%	11.4%	4.8%	6.8%	6.3%	0.8%	4.1%	1.8%	0.3%	2.5%	1.8%	1.8%	0.3%	1.0%	
Michael Finneran	38.2%	11.5%	2.7%	2.3%	10.0%	4.1%	3.2%	6.3%	3.8%	1.4%	5.4%	1.8%	0.5%	0.5%	2.7%	2.3%	3.4%	
Michael Fitzpatri..	16.0%	16.0%	8.0%	8.0%	8.0%	4.0%	8.0%	12.0%	4.0%			4.0%	4.0%		8.0%			
Michael Kennedy	22.4%	28.6%	8.3%	1.8%	13.0%	10.7%	2.8%	1.6%	0.7%	2.1%	1.2%	0.4%		0.4%	2.7%		3.4%	
Michael McGrath	19.8%	10.5%	10.6%	5.9%	9.8%	14.9%	5.4%	1.1%	2.7%	1.3%	1.5%	1.3%	1.1%	6.5%	2.3%	1.7%	2.3%	1.4%
Michael Moynihan	28.2%	10.3%	11.1%	4.7%	12.6%	3.4%	5.6%	3.2%	2.6%	3.8%	1.3%	1.7%	1.1%	1.7%	3.4%	1.1%	0.8%	3.4%
Michael Mulcahy	30.4%	18.7%	8.2%	3.5%	11.1%	11.7%	2.3%	1.2%	0.6%	0.6%	1.8%		1.8%	1.8%	0.6%	5.3%	0.6%	
Michael P. Kitt	29.5%	39.0%	3.5%	6.1%	4.1%	1.7%	1.6%	1.6%	2.1%	1.2%	0.6%	1.8%	1.0%	1.6%	0.5%	1.3%	0.1%	2.7%
Michael Smith	44.4%				22.2%		11.1%	11.1%						11.1%				
Michael Woods	48.6%	15.0%	5.6%	6.5%	2.8%	1.9%	2.8%		0.9%	6.5%	2.8%	2.8%			3.7%			
Micheál Martin	20.2%	16.6%	14.6%	7.8%	7.8%	4.8%	5.5%	1.4%	3.9%	1.7%	1.3%	2.2%	2.9%	3.4%	1.6%	1.1%	1.0%	2.1%
Ned O'Keeffe	25.3%	15.1%	6.2%	4.1%	7.5%	23.3%	5.5%	3.4%	0.7%	1.4%	1.4%		3.4%		0.7%		2.1%	
Niall Blaney	100.0%																	
Niall Collins	21.2%	14.5%	11.6%	9.6%	8.0%	4.0%	1.7%	1.5%	1.7%	3.7%	1.4%	2.7%	9.3%	4.1%	2.0%	0.7%	0.5%	1.8%
Noel Ahern	17.4%	7.8%	9.9%	4.6%	17.3%	1.8%	3.4%	5.4%	4.6%	1.8%	1.4%	0.5%	7.2%	1.4%	0.6%	14.2%	0.6%	
Noel Davern	33.3%				33.3%											33.3%		
Noel Dempsey	24.2%	19.1%	10.9%	7.8%	8.0%	2.9%	4.7%	2.5%	2.8%	3.5%	4.0%	1.2%	1.9%	0.6%	3.8%	0.5%	1.5%	
Noel O'Flynn	41.5%	35.1%	1.7%	1.1%	3.6%	1.7%	4.2%	0.8%	0.6%	1.1%	5.0%	0.8%	2.2%	0.3%	0.3%			
Noel Treacy	23.0%	4.4%	24.5%	18.6%	18.2%	3.8%	1.6%	0.6%	0.3%	0.3%	0.6%		1.3%	1.9%	0.6%		0.3%	
Ollie Wilkinson	28.6%		7.1%		42.9%		14.3%						7.1%					
Pat Carey	15.8%	35.8%	7.0%	4.4%	11.0%	2.6%	2.3%	5.3%	3.6%	3.7%	4.8%	0.7%	1.0%	0.1%	1.0%	0.8%	0.3%	
Pat Moylan		100.0%																
Pat the Cope Gall..	29.4%	11.5%	8.2%	7.1%	5.9%	0.4%	5.2%	5.9%	12.3%	2.2%	1.5%		2.2%		0.7%	7.1%	0.4%	
Peter Kelly	13.5%	16.3%	12.8%	2.8%	18.4%	6.4%	5.7%	5.0%		1.4%	2.1%	2.8%	2.8%	5.7%	2.8%		1.4%	
Peter Power	22.2%	21.9%	9.0%	4.5%	10.4%	11.5%	3.4%	1.4%	1.1%	3.4%	3.4%	0.6%	0.6%	2.0%	3.1%		1.7%	
Robert Troy	20.3%	15.2%	13.7%	7.6%	9.8%	3.4%	1.8%	2.5%	4.7%	1.7%	0.6%	8.9%	1.2%	1.2%	1.2%	2.7%	0.8%	2.8%
Rory O'Hanlon	44.5%	13.6%	8.7%	17.5%	3.9%		0.4%	1.1%	1.8%	3.7%	1.4%		2.6%	0.8%		0.1%		
Seamus Brennan	21.2%	8.5%	12.5%	8.6%	20.3%		1.2%	12.5%	4.6%	1.7%	1.7%			5.4%	1.4%	0.5%		
Seamus Kirk	18.7%	43.9%	3.2%	10.2%	2.2%	3.1%	4.6%	0.6%	0.8%	2.1%	3.9%	3.3%	0.1%	0.5%	1.8%	0.1%	0.7%	0.1%
Seán Ardagh	31.9%	25.2%	4.1%	8.1%	7.8%	4.8%	3.3%	0.7%	3.3%	0.7%	1.1%		3.0%	2.6%	1.9%	1.1%	0.4%	
Seán Connick	16.3%	17.7%	0.7%	2.1%	14.2%	6.4%	5.7%	5.0%	2.1%	3.5%	3.5%	9.2%	0.7%	0.7%	9.9%		2.1%	
Sean Fleming	25.5%	13.4%	10.0%	8.9%	11.4%	4.8%	2.5%	1.6%	4.1%	1.2%	3.0%	1.7%	1.3%	2.3%	2.9%	1.5%	1.2%	2.6%
Seán Haughey	12.1%	9.2%	6.3%	4.5%	10.6%	4.9%	2.2%	27.0%	2.9%	2.0%	4.0%	7.4%	0.9%	0.2%	4.5%		1.1%	
Seán Ó Fearghail	23.6%	13.3%	10.5%	6.8%	9.4%	1.7%	5.9%	2.7%	4.2%	3.6%	1.4%	3.2%	5.0%	1.5%	0.6%	4.4%	0.5%	1.6%
Seán Power	23.7%	12.8%	5.4%	8.3%	14.5%	1.4%	8.1%	4.7%	7.4%	5.5%	1.7%	0.2%	2.1%	0.9%	1.0%	0.7%	1.6%	
Síle de Valera	17.0%	3.8%	9.4%	3.8%	17.0%			35.8%	1.9%	7.5%				1.9%		1.9%		
Thomas Byrne	14.2%	18.5%	6.2%	5.8%	18.2%	13.1%	2.9%	5.5%	1.1%	2.9%	1.8%	1.1%	2.9%	1.1%	2.9%		1.8%	
Timmy Dooley	21.3%	19.8%	14.8%	7.3%	6.6%	4.5%	3.2%	1.9%	3.3%	1.9%	0.6%	1.9%	2.1%	3.4%	2.4%	1.9%	1.1%	2.1%
Tom Kitt	41.3%	25.5%	8.3%	4.2%	6.3%	0.3%	1.0%	2.4%	4.8%	2.0%	1.4%	0.2%	1.5%	0.3%		0.5%		
Tom McEllistrim		5.3%	5.3%	5.3%	21.1%	5.3%	5.3%	36.8%					5.3%		5.3%		5.3%	
Tony Dempsey	25.0%		25.0%		50.0%													
Tony Killeen	30.2%	14.3%	3.9%	8.3%	9.2%	3.6%	4.5%	4.1%	2.1%	4.8%	5.4%	1.3%	2.0%	1.0%	3.5%		1.9%	
Willie O'Dea	21.3%	16.5%	12.8%	9.3%	9.1%	3.7%	2.2%	1.6%	2.0%	3.5%	2.5%	2.4%	2.6%	2.9%	3.6%	1.5%	1.2%	1.1%

Figure A.8: Fianna Fáil -TD by topic coverage% of total speeches made by a TD (continued)

Appendix B

Python Code (Sample)

B.1 Data pre-processing-exract nouns code

```
#!/usr/bin/env python

import sys
import os, os.path, sys, codecs
import logging as log
from optparse import OptionParser
import text.util

#singlarize
import inflection

import nltk
import nltk.data
from nltk.tag import pos_tag
from nltk.tokenize import word_tokenize
```

```
# Parses arguments


def main():
    parser = OptionParser(usage=”usage: %prog [options] directory1 directory2 ...”)
    # Parse command line arguments
    (options, args) = parser.parse_args()
    if( len(args) < 1 ):
        parser.error( ”Must specify at least one directory” )
    log.basicConfig(level=20, format=’%(message)s’)

    # Process each directory
    for in_path in args:
        dir_name = os.path.basename( in_path )
        for filename in os.listdir (in_path):
            filepath = os.path.join(in_path, filename)
            with open (filepath, ”r”) as file :
                text=file.read().replace(’\n’, ’ ’).strip()



                # Each sentence is then tokenized and POS tagged.


                sent_detector = nltk.data.load(’nltk:tokenizers/punkt/english.pickle’)
                for sentence in sent_detector.tokenize(text):
                    tokenizedSentence = word_tokenize(sentence)
                    taggedSentence = pos_tag(tokenizedSentence)
                    start = True
                    currentCandidate = []


                    for word, pos in taggedSentence:
                        if start:
                            start = False
                            continue
                                        # Identify singular and plural nouns and proper
                                            nouns
                        if (pos == 'NN' or pos == 'NNP' or pos == 'NNS' or pos ==
                             'NNPS'):
                            currentCandidate.append(inflection.singularize(word.lower()))
                            continue

                    out_dir = "data/sample"
                    output_filepath = os.path.join(out_dir, dir_name, filename)
                    output_text = "\n".join(currentCandidate)
                    with open(output_filepath, "w") as file:
                        file.write(output_text)
        print(dir_name)
#--------------------------------------------------------------------------------------------

if __name__ == "__main__":
    main()
```