National College of Ireland

Topic Analysis for House Rent in Ireland

MSc Research Project
Data Analytics

Oluwasurefunmi Maria Idowu
Student ID: x18158188

School of Computing
National College of Ireland

Supervisor: Noel Cosgrave

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Oluwasurefunmi Maria Idowu
Student ID:	x18158188
Programme:	Data Analytics
Year:	2019
Module:	MSc Research Project
Supervisor:	Noel Cosgrave
Submission Due Date:	12/12/2019
Project Title:	Sentiment Analysis and Topic Modelling for House Rent in Ireland
Word Count:	XXX
Page Count:	14

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	7th December 2019

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Contents

Sentiment Analysis and Topic Modelling for House Rent in Ireland

Oluwasurefunmi Maria Idowu
x18158188

Abstract

Being in the era of social media, so many platforms have been created to enable users express themselves in different ways. These posts by users have produced useful information to a wide variety of business owners and organizations thereby improving their decision making process. This paper presents a CRISP-DM methodology for text mining of house rent data in Ireland using unsupervised machine learning techniques (Latent Dirichlet Allocation), as well as supervised machine learning techniques (logistic regression, SVM and Gradient Boosting). The solution involves the use of R language to parse web posts about rent within the last year from a popular Irish website known as boards.ie, thereby creating useful insights. Before classification. The mean sentiment scores per month were aggregated and tested using ANOVA and the results show that the sentiments within each month are not statistically different from each other, which infers that is no need to work on the topic modelling per month. Furthermore, the texts were classified into positive and negative texts and topic modelling was done for each class in order to distinguish between positive and negative topics. The classifier results show that generalized linear models outperformed linear SVM with an accuracy of 74.69% and a t-test was performed to conclude that there was a significant difference in the results. In conclusion, the work aims to contribute to the complex resolution of accommodation issue in Ireland, leaving room for more research in the future.

Keywords topic modelling, data mining, natural language processing, web scraping, machine learning, text mining, housing, rent

1 Introduction

1.1 Background and Motivation

The progression from web 1.0 to web 3.0 has created room for users to interact freely using different technologies and applications. It is safe to say that the purpose has been fulfilled successfully, as the amount of data created by internet users have increased exponentially in the last decade. This intense growth has made the web become the biggest readily available and easy to use source of data in the world (Bing; 2011; Russell and Klassen; 2019). **(quote something about russell)** This kind of data is known as unstructured data and accounts for emails, call and speech transcripts (which happens when calls and speeches are recorded), social media outreach (Twitter, Facebook and YouTube, etc), field agents, sales people, interviews ans surveys. Although this paper strictly focuses on

ToDo

textual unstructured data, some other forms of non-textual unstructured data exist in the form of images and videos. In general, unstructured data covers 80% of all forms of data in the world, and text mining is the part of data science which deals with unstructured data (Kotu and Deshpande; 2019). It initially began to surface when the need to sort and arrange multiple documents arose (Cutting; 1992). The evolution of this information age has come with so many jobs and solutions. Firms have seized the opportunity brought about by text mining for their benefit by understanding customers' perspectives about their products and services.

Topic modelling is a hot topic in text mining, as it deals with the clustering of several documents together to find reoccurring patterns. It has been used in several domains such as government services (Aziz et al.; 2018), music (Laitonjam et al.; 2015). Furthermore, Beykikhoshk et al. (2018) used biomedical data, Brookes and McEnery (2019) sourced patient feedback, and Nikolenko et al. (2017) gathered several blog posts in Russian language. The dynamic data selection from different domains has made the research field even more interesting. So in this research, rent data is used. More details will be discussed in the following sections.

1.2 Significance of the Research

Accommodation has always been a predominant topic in Ireland, especially in the city of Dublin, and according to the minister of housing Murphy and McCarthy (2017), it is quite a difficult to resolve to resolve. Nonetheless, he further states in the 2020 strategy that contributions to this complex solution are welcome. Hence, carrying out research in this area is highly significant. Moreso, there have been solutions to predicting house prices with the use of supervised machine learning techniques (Varma and Nair; 2018; Mukhlishin and Wibowo; 2017), as well as sentiment analysis solutions (Yang and Seng; 2017; Kou et al.; 2018). However, there is a gap in the domain of research has been a gap as the use of unsupervised learning algorithms, have not been explored on rent data.

Therefore, the aim of this project is summarized into research questions and objectives as follows.

1.2.1 Research Question

Q1: "Is there a significant difference between the average sentiment polarity scores per month within the last year?
Q2a: "How well can an unsupervised machine learning technique such as Latent Discriminant Analysis produce clusters of predominant topics revolving around rent in Ireland have been discussed within the last 12 months?"

Q2b: What underlying topics tend towards the positive, neutral and negative posts?
Q3a: How well can supervised machine learning algorithms such as logistic regression and Support Vector Machine classify the sentiment polarity of each post?

Q3b: Is there a significant difference in the performance of these two algorithms?

1.2.2 Objectives

In order to effectively answer the research question(s), the following objectives must be achieved:

- A critical analysis of previous works done on housing, topic modelling and text classification techniques

- Web scrapping: Extracting all posts about rent within November 2018 and 18th November 2019, from the Accommodation and Property Forum on boards.ie
- Sentiment analysis: Performing a sentiment analysis on all posts and testing if monthly sentiments are significantly different.
- Topic Modelling: Perform LDA on all texts and sentiment terms to discover the prevalent topics discussed on the forum within the last one year
- Classifying sentiments based on sentiment terms
- Implementing classification models such General Linear Models and Support Vector Machines and analysing the significance of the results

As far as the research objectives are concerned, the only limitation to reaching the goal of this project would be the memory size of the tool used (R Studio). However, it maintains the assumption that each post... **(write assumptions)**. This work hopes to advance the Data Analytics field with text mining solutions pertaining to housing, which will also be useful to the housing agency, as well as the department of housing, planning and local government in Ireland (Davidson; 2017; Murphy and McCarthy; 2017). **ToDo**

The order of this paper is as follows. The next chapter gives a critical review of previous works done in this field to justify the research. The following section outlines the required method and procedure. The next section discusses the tools and techniques used for implementation, as well as the architectural design, which is followed by implementation in chapter 4, evaluation of results in chapter 5, and research findings and challenges are discussion in chapter 6. Finally, in chapter 7, all conclusions are drawn, as well as future research are recommended.

2 Literature Review

2.1 Machine Learning and Housing

One of the basic needs that enable human beings operate properly is shelter. We can all attest to that great feeling of going home to our beds and having a good night rest. It is mentally relaxing. Little wonder then, that Maslow (1943) considered it to be one of the first components in the classic theory, explaining the hierarchy of human needs. This validates continuous research for solutions related to housing.

Giving the constant demand for shelter, researchers have gone far and wide in creating solutions for this need. One of which is the price of a house. This is so important, as one of the reasons for homelessness is poverty **(cite)**. A lot of authors have come through with solutions by exploring the correlation between house prices and certain variables. Varma and Nair (2018) took into consideration the proximity of accessible facilities and services such as hospital, supermarket, park, and train station, using Google Maps API, while (Mukhlishin and Wibowo; 2017) gave more attention to the tax on land, as well as the physical state and age of the house. Although both authors used Neural Networks, Varma and Nair (2018) focused more on building a concrete regression models by implementing linear regression, random forest regression (using bag of trees), and Gradient Boosting, while Mukhlishin and Wibowo (2017) implemented fuzzy logic, and K-nearest neighbors. **ToDo**

Furthermore, some authors adopted the use of unstructured data to carry out meaningful research in this area (Yang and Seng; 2017) (Kou et al.; 2018). These papers

were found to include similar research, having the same goal - to analyze the correlation between real estate prices and reports from several online newspapers, and using the same technique- sentiment analysis.

2.2 Topic Modelling as a Text Classification Technique

There are several mining techchiques in text analytics, but this research covers only two - topic modelling and text classification.

2.2.1 Topic Modelling

Topic Modelling involves finding groups of words (called topics) within a large collection of documents(Kotu and Deshpande; 2019). It was developed by David The general purpose is for exploratory analysis, to find patterns and create knowledge from the output. It's application in existing fields and industries is quite wide, from medicine (Beykikhoshk et al.; 2018) to music (Laitonjam et al.; 2015), and then to government services (Aziz et al.; 2018). It is ofetn disovered that the longer the text, the better the model

Some do it alone, while others like Aziz et al. (2018) combined it with text classification to yield more insights. In the paper (Aziz et al.; 2018), topic modelling was used to identify dominant topics related to government service satisfaction. Furthermore, text classification was used to classify the polarity of the texts and topic modelling was used to distinguish between negative and positive topics revolving around the subject. Although the data was not large, the results were pretty impressive, with an accuracy of over 80%, SVM being the classifier. Evaluation was done using precision accuracy and F-score and no of topics were gotten using coherence score.

Beykikhoshk et al. (2018) used topic modelling over time (TOT) to compare changes in topics over time

Brookes and McEnery (2019) gathered feedback of the National Health Service patients, posted on their online platform. The large dataset consisted of over 200,000 comments within early 2013 and late 2015. In the first phase of the research, it was discovered that relying on only topic word lists was not so effective for defining topics. In the second phase, some reoccurring conjunctions were taken into consideration, implying that the grouped words may mean something else, for example, a mixture of positive and negative feedback. Furthermore, the deletion of punctuation marks tends to have an effect on the outcome of the model.

Laitonjam et al. (2015) used hierarchical LDA and incorporated time into it, creating a dynamic HLDA-TOT model. A portion of the popular Million Song Dataset (MSD) was used, consisting of 1000 songs within 1990 and 2010, and using 100 songs per year. It was discovered that the change in topic frequency was found, in addition to the change in topic. It was also mentioned that HLDA has a parameter for determining the number of topics, while DTM does not. The only limitation was that HLDA didn't really have a structured formula for Gibbs sampling, unlike LDA. HLDA-TOT performed similarly to HLDA when the held-out log likelihood was used as the evaluation metrics.

LDA is a kind of topic

Therefore, this paper will only discuss some of the previous works which are relevant to the research.

Often referred to as probabiistic models,

These words are called topics and the documents are

Paper	Domain	Data Source	Model	Evaluation Metrics
Aziz et al. (2018)	Government service satisfaction	Twitter	LDA	.....
Beykikhoshk et al. (2018)	Biomedical data	Abstracts of scientific papers	Topic over tome (TOT)	....
Brookes and McEnery (2019)	Patient feedback	NHS choices online services	LDA using Mallet	Topic allocation - 71.37%
Laitonjam et al. (2015)	Music	Million Songs Dataset	HLDA, LDA and HLDA-TOT	Topic Popularity
Nikolenko et al. (2017)	Entertainment	Russian language blogs	...	....
Schmiedel et al. (2019)	Structured	Online reviews	....	.....
Cool (2018)	Structured	....	....	.....
Schmiedel et al. (2019)	Structured	.....	....	.....
Schmiedel et al. (2019)	Structured	.....	....	.....
Schmiedel et al. (2019)	Structured	.....	....	.....

Table 1: A literature summary of topic modelling applications.

We could conclude that topic modelling is a simple, quick and easy solution, while topic classification takes a longer time to solve which will promote business growth.

In Table 3 Literature Summary of Topic Modelling Applications.

2.2.2 Model Evaluation and Interpretation

2.3 The Irish Housing Industry in Retrospect

There are two critical bodies that play significant roles in the housing industry of Ireland. The first is the housing Agency of Ireland [1]. This body takes responsibility for the welfare of the community by constantly being acquainted with their housing related needs and wants, and finding ways to resolve those needs. On the other hand, the Department of Housing, Planning and Local Government operates at higher level, by producing official documentaries about the rate of homelessness within the nation, as well as other related statistics. They are also working in conjunction with private, public, and voluntary bodies to create better solutions for the housing issue in the country. One of which is to

[1] `www.housingagency.ie`

optimize the standard of "rented housing" and reinforce the protection of tenants and landlords [2]. Although, according to reports show that the average rate of homelessness remains within the range of 6000 plus, the minister of housing, planning and local government, Minister Eoghan Murphy, earlier mentioned in the three-year strategy that the accommodation crisis in Ireland is worth contributing to, regardless of how difficult the solution is (Murphy and McCarthy; 2017). Two important terms were predominant in the findings; Residential Tenancies Board Ireland (RTB) [3] and Housing Assistance Payment (HAP) Scheme [4]. They represent the **(talk about them)** **ToDo**

2.4 Conclusion

3 Methodology

Before delving into the details of the research procedure, the following tools and techniques R Studio for coding and visualization
Tableau for some of the visualization
Techniques: SVM, Logistic Regression

3.1 About the Data

The data was extracted from a well known Irish website - boards.ie [5]. Its main purpose is quite simple; to create room for discussion of numerous topics in Ireland. Ever since its establishment in 2000, it has remained relevant and successful with over 60 million posts in their database, and over 8,000 posts daily. The platform is rich in information based on several topics which are usually discussed, ranging from politics to entertainment, business and finance, home, interior design and lots more.It also has several forums like Taxation, Accommodation and Property, Dublin City, and so on, which make it much easier for users to find the information they are looking for. With the above description and facts, it can be concluded that there is no better place to find unstructured data about hot topics in Ireland. Another reason for this choice is its uniqueness as it has never been used in this way. The last published article which used over 10 years ago. It was a collaboration by the research department in NUI Galway and the software team in IBM Dublin Chan et al. (2017).

the data contained a mean length of....

3.1.1 Ethics

Permission to extract the web data was requested from boards.ie administration, using the R user client **(add user client)** and granted. An ethics application form was filled, **ToDo** and a declaration form was signed.

3.2 Data Extraction and Pre-processing Pipeline

One of the most popular methods of scraping web-based data for text mining and analysis is the use of the rvest package provided by in R studio **(cite a paper)**. All posts **ToDo**

[2] https://www.housing.gov.ie/
[3] rtb.ie
[4] hap.ie
[5] www.boards.ie

related to rent were scraped from the Accommodation Property forum using the keyword "rent". The decision was made after screening through posts from the other forums and realizing that they were neither useful nor relevant to the research. Furthermore, the initial plan was to harvest all posts within the last 10 years, but due to the high volume of data available, it was concluded that only posts within the last year would be used, resulting to a critical analysis of posts within the last 12 months. During the extraction period, all posts about rent within the last 14 years were consist of Over 5600 pages, whereas the posts within the last year consist of exactly 450 pages, returning exactly 6750 observations. The said observations were converted to a csv file for easy accessibility.

3.3 Modelling

3.4 Evaluation Metrics

Well, you already know that supervised machine learning models need to be fed training data that is relevant to your predefined topic list. By using a subset of that tagged data to test your model, you can find out how accurate it is at making predictions. You can test the actual tag for a specific text and compare it to the predicted tag then, with the results, calculate the following evaluation metrics:

Accuracy: the percentage of texts that were assigned the correct topic Precision: the percentage of texts the classifier tagged correctly out of the total number of texts it predicted for each topic Recall: the percentage of texts the model predicted for each topic out of the total number of texts it should have predicted for that topic F1 Score: the average of both precision and recall The best way to test your model and receive these evaluation metrics is through cross-validation. This involves randomly splitting your training dataset randomly into equally sizes (e.g 4 sets, each with 25% of the data), then training a classifier with one of these sets, and using the remaining unseen 75% to test your classifier. Then you can build the final model by using all the sets as training data. **(paraphrase!)** **ToDo**

4 Design and Implementation

The Cross-industry Standard Process for Data Mining (CRISP-DM) initiated by Chapman et al. (2000) is the chosen method for this research.

A 3-tier design consisting of the data layer, business layer and client layer is used as a framework for the project. The process flow of this project is presented in Figure 1, which includes the names of all tools and techniques used

The techniques and/or architecture and/or framework that underlie the implementation and the associated requirements are identified and presented in this section. If a new algorithm or model is proposed, a word based description of the algorithm/model functionality should be included. You will of course want to discuss the implementation of the proposed solution. Only the final stage of the implementation should be described.

It should describe the outputs produced, e.g. transformed data, code written, models developed, questionnaires administered. The description should also include what tools and languages you used to produce the outputs. This section must not contain code listing or user manual description.

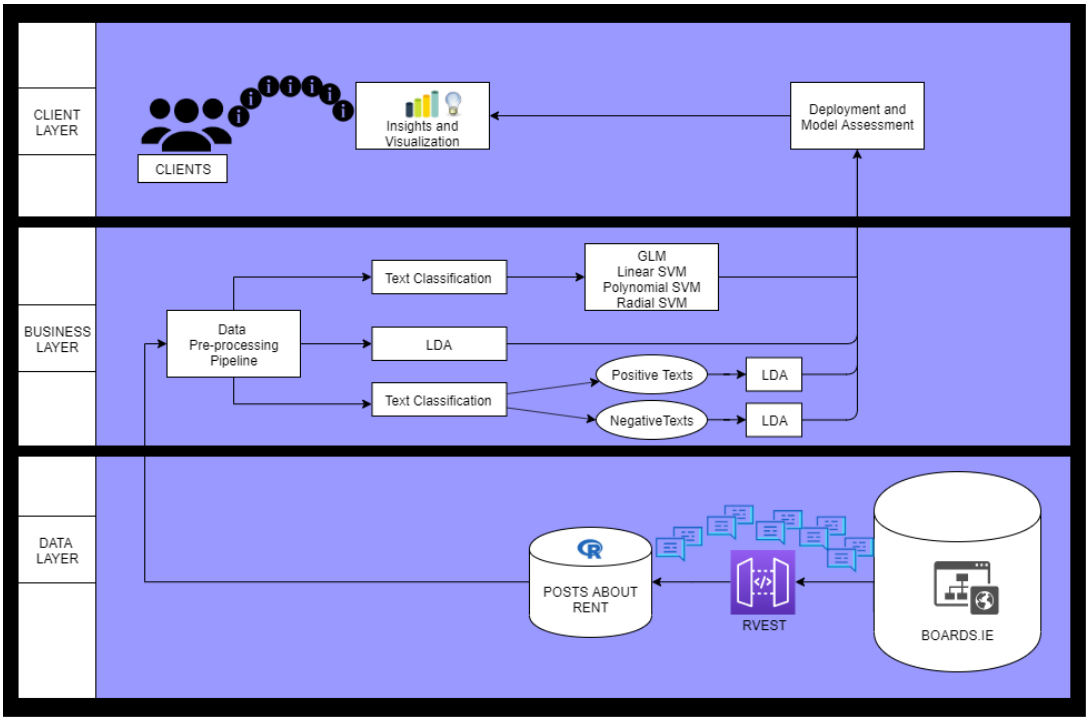

Figure 1: A framework for the research implementation

4.1 Sentiment Analysis

Sentiment analysis was performed using R's sentimentr package to calculate sentiment polarity score of each text. Two experiments were performed here.

- Sentiment Scores with Punctuation Marks: Calculated scores without breaking the sentences
- Sentiment Scores without Punctuation Marks: Calculates scores by breaking texts the sentences (element_id) and calculating the polarity of each sentence. Furthermore, the mean sentiment scores

Sentiment with punctuation produced more concise results than the former, as the range was smaller than that of the former.

4.1.1 Mean Sentiment Scores per Month

5 Evaluation

5.1 Exploratory Text Analysis

One major way of visualizing text data is the use of WordCloud. In R, wordcloud and wordcloud2 packages were used to show the most frequently used words in the text as well as the most

Figure 2,

Figure 2: Word Cloud

5.2 Sentiment Analysis

The ANOVA test is used to verify if there is a statistical significant difference between the mean sentiments over the last 13 months, that is, November 2018 to November 2019. The null and alternative hypothesis are stated below:

Null Hypothesis: No sentiment score is statistically different is from the others

Alternative hypothesis: At least one sentiment score is significantly different from the others

The output below records an F value of .0279 which indicates that we do not reject the null hypothesis. This result leads to the conclusion that there is no need to partition the data into months, it is better to carry out the analysis as a whole document.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
monthly.sentiments$mean.sentiment	1	4.49	4.495	0.279	0.608
Residuals	11	177.51	16.137		

5.2.1 Categorizing the Sentiment Scores

Afterwards, the sentiments polarity scores were grouped into positive and negative sentiments. The polarity score is calculated using combination of version of Joacker's (2017) and Rinker's dictionaries (Hu and Liu; 2004).

5.3 Topic Evaluation

There are several ways of evaluating topic models, which causes a disunion in the field (Blei; 2012). One can train the modelled data, then test for similarity, while one can simply use all the data, since it is an unsupervised technique (Kotu and Deshpande; 2019). This research only uses the splitting technique for the text classification phase.

Just like k-means clustering, the number of topics (k) must not always be determined by an algorithm, especially when there business objective in place. Similarly, k is chosen as 10 **(check this)** to fit the purpose of this thesis. Implementing HDP might affect the desired results. **ToDo**

In LSA, the distribution has no assumptions which tends to result in an obscure model. Alternatively, LDA assumes a Dirichlet distribution for words to topics, and topics to documents. Furthermore there are two hyper parameters that control document and topic similarity, known as alpha and beta, respectively.
ALPHA: This is the measure of the number of topics. A high alpha indicates that every document is likely to contain a mixture of most of the topics, while a low alpha means that a document is likely to be a representation of just a few of the topics.
BETA: This is measure of the number of words A low value of alpha will assign fewer topics to each document whereas a high value of alpha will have the opposite effect. A low value of beta will use fewer words to model a topic whereas a high value will use more words, thus making topics more similar between them.
GAMMA: The probability of a particular topic being in a document
ENTROPY: This is the measure of randomness. A high entropy indicates high randomness, which is not good, but low entropy means the level of randomness is low, which is usually what we want.

There is another hyperparameter called the coherence probability, which is used to determine the optimum number of topics because the LDA algorithm does not include this calculation.

5.3.1 The Topics

As reccommended by Sievert and Shirley (2014) Figure 3 shows all topics one to ten, with their prevalent words.

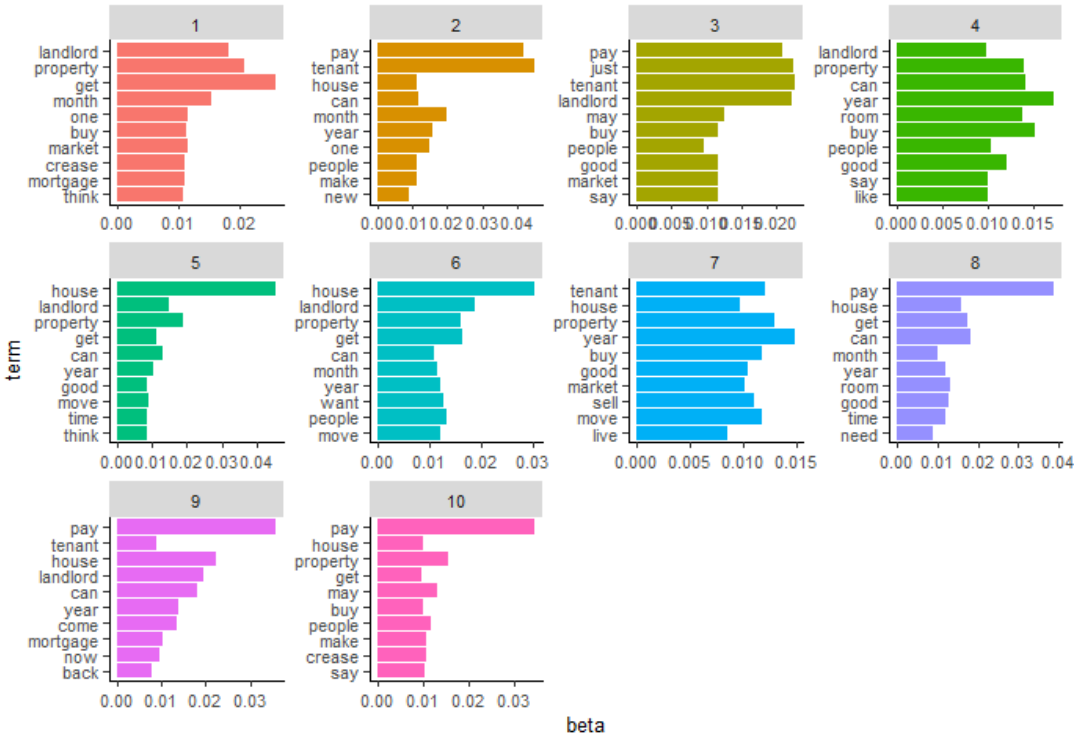

Figure 3: Visualization of the topics

5.3.2 LDA Model Evaluation

Results are provided in Table 3.

Table 2: A table caption.

	Fit	Fixed	Gibbs	CTM Fit
Alpha	34.67176		13.65	
Entropy	2.302519	2.300400	2.268220	
Gamma	stuffed		92.50	
Likelihood	stuffed		33.33	

5.4 Text Classification Model Results

The following performance metrics are used to classify the texts appropriately to get optimum results. **(cite relevant literature)** **ToDo**

Precision: This is the degree to which the positive cases are indeed positive. It is the extent to which positive cases were predicted correctly (Lantz; 2013). In addition, the precision formula is defined thus:

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

Recall: On the other hand, recall is the number of cases which should have been positive but were not declared positive Lantz (2013). The formula for recall is presented thus:

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

F-measure: Also known as F-score, basically measures how well each model performs. It is the combination of the precision and recall value, ranging from 0 to 1. It explains the strength and weakness of each model. For instance, if the F-measure of our SVM model is 0.9, it is considered very strong, while 0.1 is considered very weak. The formula for F-measure is therefore defined as follows:

$$F - measure = \frac{Precision * Recall}{Precision + Recall} \tag{3}$$

In Table 3, the results are presented according to Lantz (2013) standards.

Table 3: Text Classifier Results in Percentage

Classifier	Precision	Recall	Bal.Accuracy	Duration
GLM	74.51	76.99	75.73	2.7 mins
Linear SVM	74.28	74.38	73.69	1.17 hrs
Radial SVM	73.08	76.99	76.32	21.9 mins

5.5 Experiment / Case Study N

...

6 Discussion

A detailed discussion of the findings from the N experiments / case studies. Note that this discussion will have a lot more detail than the discussion in the following section (Conclusion). You should criticize the experiment(s), and be honest about whether your design was good enough. Suggest any modifications and improvements that could be made to the design to improve the results. You should always put your findings into the context of the previous research that you found during your literature review

- The most tedious part of the implemnation was the cleaning aspect
- deciding k - no of topics
- limitations
- was it good enough?
- why was glmnet better?
- modifications and improvements

7 Conclusion and Future Work

This paper examines the topics discussed about rent within the last one year. It further experimented on topics with negative sentiment polarity and topics with positive sentiment polarity. The research question... The results show that.... The key findings are... The Efficacy of this research.. The limitations..

However, there are diverse approaches in analyzing the given data. Due to high volume of data on the forum, only posts within the last year were extracted. In the future, posts from the last 10 years can be extracted, as well as posts from other forums that represent other locations in Ireland, such as Dublin city, Cork, Limerick and so on. Also, further techniques such as aspect based sentiment analysis can be carried out to get deeper insights for the text. Topic classification..

Regardless of the limitations in this paper, the aim of this work was successfully achieved, as its general purpose was to water the ground for more analysis on rent data in the future.

References

Aziz, M. N., Firmanto, A., Fajrin, A. M. and Hari Ginardi, R. V. (2018). Sentiment analysis and topic modelling for identification of government service satisfaction, *2018 5th International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE)* pp. 125–130.

Beykikhoshk, A., Arandjelović, O., Phung, D. and Venkatesh, S. (2018). Discovering topic structures of a temporally evolving document corpus, **55**(3): 599–632.

Bing, L. (2011). Web data mining: exploring hyperlinks, contents, and usage data, *New York: Springer, 2nd ed* pp. 353–358.

Blei, D. M. (2012). Probabilistic topic models, **55**(4): 77.

Brookes, G. and McEnery, T. (2019). The utility of topic modelling for discourse studies: A critical evaluation, **21**.

Chan, J., Hayes, C. and Daly, M., E. (2017). Decomposing discussion forums and boards using user roles, *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media* pp. 215–218.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. and Wirth, R. (2000). CRISP-DM 1.0, *Step-by-step data mining guide* .

Cool, M. (2018). Topic modelling: Location-based offline advertising using twitter.

Cutting, D. K. (1992). Scatter/gather: A cluster-based approach to browsing large document collections, *Copenhagen proceedings of the 15th annual international ACM SIGIR conference on research and development in information retrieval* pp. 318—-329.

Davidson, N. M. (2017). Affordable housing law and policy in an era of big data, *Fordham Urban Law Journal* **44**(2): 277—-300.

Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews, *Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining* pp. 1168–177.

Kotu, V. and Deshpande, B. (2019). Chapter 9 - text mining, *Data Science (Second Edition)* pp. 281–305.

Kou, J., Fu, X., Du, J., Wang, H. and Zhang, G. Z. (2018). Understanding housing market behaviour from a microscopic perspective, *2018 27th International Conference on Computer Communication and Networks (ICCCN)* **115**: 1–9.

Laitonjam, N., Padmanabhan, V., Pujari, A. K. and Lal, R. P. (2015). Topic modelling for songs, pp. 130–135.

Lantz, B. (2013). Machine learning with r:learn how to use r to apply powerful machine learning methods and gain an insight into real-wold applications, *Birmingham: Packt Publishing Ltd.* .

Maslow, A. H. (1943). A theory of human motivation, *Psychological Review* **50**(4): 370–396.

Mukhlishin, M. F., S. R. and Wibowo, A. (2017). Predicting house sale price using fuzzy logic, artificial neural network and k-nearest neighbor, *2017 1st International Conference on Informatics and Computational Sciences (ICICOS),* pp. 171–176.

Murphy, E. T. D. and McCarthy, J. (2017). Department of housing, planning and local government: Statement of strategy 2017-2020, pp. 5–43.

Nikolenko, S. I., Koltcov, S. and Koltsova, O. (2017). Topic modelling for qualitative studies, **43**(1): 88–102.

Russell, A. M. and Klassen, M. (2019). Mining the social web: Data mining facebook, twitter, linkedin, instagram, github, and more, *Canada: O'Reilly Media Inc. 3rd edition* .

Schmiedel, T., Müller, O. and vom Brocke, J. (2019). Topic modeling as a strategy of inquiry in organizational research: A tutorial with an application example on organizational culture, **22**(4): 941–968.

Sievert, C. and Shirley, K. (2014). LDAvis: A method for visualizing and interpreting topics, *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, Association for Computational Linguistics, pp. 63–70.

Varma, A., S. A. D. S. and Nair, R. (2018). House price prediction using machine learning and neural networks, *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)* pp. 1936–1939.

Yang, H.-F. and Seng, J.-L. (2017). Using sentiment analysis to explore the association between news and housing prices, *Asian Conference on Intelligent Information and Database Systems* pp. 170–179.