National College of Ireland

Topic modelling: Location-based offline advertising using Twitter

MSc Research Project
Data Analytics

Martijn Cool
x16101529

School of Computing
National College of Ireland

Supervisors: Dr. Paul Stynes
Dr. Dympna O'Sullivan
Dr. Pramod Pathak

Topic modelling: Location-based offline advertising using Twitter

Martijn Cool
x16101529
MSc Research Project in Data Analytics

9th August 2018

Abstract

Targeting the desirable audience for offline advertisements may be challenging. Unlike online where user profiling is common practice, offline advertising can only target a broader, less appropriate audience. Offline is generally more expensive than online marketing, but your advertisements may be overshadowed with the greater amounts of online advertising content, therefore, offline advertising is still valuable. This study researches an alternative approach to offline advertising with the use of online publicly available information from Twitter. The aim of this research is to investigate to what extent interest topics can be identified from Twitter content in a geographical region. Tweets located in Dublin County were collected and used to identify specific interest zones across Dublin. The main method for topic modelling in this research is with a Latent Dirichlet Allocation algorithm. The algorithm runs into obstacles on smaller sized documents to produce consistent, coherent and non-subjective topic labels. It is evident from the research that topic modelling on tweets may not produce consistent topic categories, to represent the Dublin population and interest zoning adequately. The research provides a well-designed framework from which future work in interest zoning using topic modelling can be done. The paper proposes additional measures to improve on location-based offline advertising using online content.

1 Introduction

The current approach to outdoor advertising is narrow-minded. When a company wants to make awareness for their brand or a new product, it is common to use billboards as a medium. A billboard is rented from a billboard company who will offer available spaces around the city. Such marketing campaigns frequently have the same billboard advertisements spread across town. Larger marketing budgets will also allow for the preference of targeting billboards by location, for example well-off areas may see different types of products or brands advertised as opposed to others. Location-based advertising is most certainly common and is nothing new (Bruner and Kumar; 2007). What is left out of the equation is what really goes on in different areas in a city; what citizens are generally talking about. Therefore, the billboards may not represent the interests of the audience that will view the billboards. One way of finding this out, however, is to use an

online medium, such as social media. Social media is increasingly being used for people to express their opinions or what goes on in their lives. Twitter is a prime example of such a platform, where the content of the posts can have valuable insights when the data are aggregated.

Twitter started in 2006 and has since grown to be one of the largest social networks in the world (*Annual Report*; 2018). Twitter allows 280 characters per post [1]. With this limit, users express themselves in what is called a tweet. Many users may do this several times in a day. Therefore, it can be assumed that users will share the same interests on Twitter as they would offline. Furthermore, each tweet is posted from a specific place and Twitter keeps track of the coordinates. However, as it is off by default, users are entitled to opt in for the tracking. [2]. Using the insights on the coordinates and tweet content, it is possible to discover patterns of what content is posted at certain locations.

This research will merge this publicly available knowledge from Twitter to integrate it as a tool for location-based advertising. Furthermore, this research will borrow techniques used for online advertising, e.g. on social media, and apply it for offline advertising. It can be assumed that social networks can represent a city's population to an extent. In this study, online social content will be analysed using Natural Language Processing (NLP) techniques. The aim of this research is to investigate to what extent interest topics can be identified from twitter content in a geographical region. This research builds upon previous work by Anagnostopoulos et al. (2017). They analysed Twitter *Lists* as a way to profile users with the intent of identifying interest zones within the city of Milan. One of the disadvantages of this approach is the scalability because user networks can be complex and will require a lot of processing. Topic modelling, on the other hand, requires less computer power since the focus is on the actual content of posts, rather than the users and the Twitter *Lists*. The major contribution of the research is to extend on marketing practices by making those more location precise to a specific audience. The hypothesis of this research is that through the content of Twitter it will be able to identify specific interest zones in Dublin.

This paper follows a specific outline. Firstly, Section 2 will cover relevant research that has been done in the field. Section 3 covers the methodology, including detailing the data. Section 4 focuses on the Design and Implementation. In Section 5, the results are presented, and later, discussed. The paper is concluded in Section 6.

2 Related Work

This section will cover the related work to the research. Due to the extensive literature coverage in the NLP field, it will be primarily focused on topic modelling, the evaluation of the techniques and location-based classification.

2.1 Information Retrieval

This research is heavily reliant on the retrieval of information. The information is extracted from the collected tweets' data. As Manning et al. (2008) put it, information retrieval can be defined as "finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections" (p.5). The

[1] Giving you more characters to express yourself: `https://bit.ly/2fQ2b7W`

[2] Tweet location FAQs: `https://help.twitter.com/en/safety-and-security/tweet-location-settings`

textual contents of the documents, which are the tweets, have valuable information when used as a collective. The information need is the identification of interest zones based on the texts.

In a recent paper, Anagnostopoulos et al. (2017) characterised Twitter users' interests to derive zones across Milan with a specific category. Their methodology comprised of two main parts; inferring user interest and top-k zone ranking. This research follows a similar approach of data collection of geotagged tweets, except Anagnostopoulos et al. had access to Twitter Firehose, which allows for enhanced parameter specifications with the API. Once collected, they used Twitter *Lists* to apply an interest on each collected user. A *Twitter List* is a feature that "allows users to create and manage curated lists of other users" and *Lists* are given a name and description (Anagnostopoulos et al.; 2017, p.528). For example, a *List* called *Rock Music* can consist of artists' accounts. Using the most common *Lists*' names, they produced nine distinct interest categories, such as *Music*, *Home* and *Sport*. Then, they assigned a category to each user based on their association with a *List*, which is now considered an interest. For the top-k zone ranking, they utilised the users coordinates to approximate an interest category for each zone. Thus, they used the users' interests from the *Lists* to identify the zones where the geotagged tweets came from. Additionally, they had access to mobile usage data to select the busiest zones. One of the disadvantages of the *List* approach is that it is not a scalable solution for larger data as the interest derivation requires extensive storage and processing to obtain data from all followed users that are related to *Lists* (Sorella; 2018).

By using influences from NLP, this paper extends on Anagnostopoulos et al.'s (2017) work by swapping *Lists* for focusing on the text to extract interest topics. This addresses the unscalable aspect of *Lists*.

2.2 Topic Modelling

This paper focuses on the modelling of topics from the text. Tweets' text have a pool of information. As Chen et al. (2010, p.1186) said "users are not passive consumers of content in information streams. People are often content producers as well as consumers".

Topic modelling has previously been applied on Twitter data. Nugroho et al. (2015) identified topics in tweets. They found that topic modelling is best for re-occurrence of similar words, which can be argued to be the case for tweets. Furthermore, they argued that topic modelling using tweets is difficult because of the limited size. However, their research relied on a dataset with tweets having a limit of 140 characters, whereas this limit has now been doubled [3]. Nonetheless, not all tweets will reach this limit. Moreover, geotagged tweets are also scarce. Cheng et al. (2010) (2014) proposed a framework to estimate the location of a tweet simply based on its content, however, this was with 160km accuracy, which would not suffice for city analysis.

The most comprehensive analysis of topic modelling algorithms was arguably done by Liu et al. (2016), who created a comprehensive overview. It is a collection of topic modelling algorithms, assessing each one. Liu et al. (2016) state most topic model algorithms are derived from the original Latent Dirichlet Allocation.

[3]Giving you more characters to express yourself: `https://bit.ly/2fQ2b7W`

2.2.1 Latent Dirichlet Allocation

The main statistical procedure that takes place is regarding the topic modelling. A Latent Dirichlet Allocation (LDA) algorithm is used for the modelling of the interest topics. LDA is a Gibbs sampling technique that was originally proposed by Blei et al. (2003). It is a rather simplistic probabilistic model and has since been redesigned. Nonetheless, the basis of the algorithm still works the same. A dataset will consist of documents, from which a corpus is created. Each document (or tweet) is likely to have multiple topics. So, the algorithm assigns a random topic to each term in the document so that a document consists of several topics. The number of topics is set beforehand. Then, the model refines its topic choices by going through each individual word. The probabilities $P(topic|doc)$ and $P(word|topic)$ are derived from the above steps. Using those probabilities, new topics are then assigned to words to increase the probabilities. Finally, the last step reiterates this approach several times depending. With Gensim, it is possible to explicitly state the number of passes that the algorithm should go through the documents (Rehurek and Sojka; 2010).

Since 2003, similar algorithms have been proposed based on the original LDA, most prominent is Hoffman and Blei (2010). In their research, they extended the algorithm by adapting it to allow for stream processing instead of batch. This enables faster processing of larger datasets. They evaluated the model on a corpus of 3 million Wikipedia pages, which has become the *Hello World* of topic modelling. Nonetheless, this research utilises a version of the original LDA, as the incentive is to use an established and proven to have worked in the field algorithm and applying it in a novel application. As the batch processing will suffice on this dataset, there is no need for online stream processing. However, this could be future work.

2.3 Evaluation of Topic Modelling

Machine learning models are generally evaluated using a type of error function. However, this can be difficult for unsupervised learning, because unlike supervised, there are no labelled data to test. Perplexity is a common measure to evaluation topics. However, this can considered to be an unreliable metric. For instance, Chang et al. (2009) found that "models which achieve better predictive perplexity often have less interpretable latent spaces" (p.2). Measures such as perplexity have issues explaining that which topic modelling aims to fulfil. Unlike perplexity, coherence measures consider the probability of two words occurring with each other. Perplexity only looks at one term at a time. Coherence measures aim to capture the context of the word to document relationships. Due to this, perplexity was not used to evaluate the topics in this research. In 2015, Roder et al. looked at many coherence metrics, including ones from scientific philosophy. Roder et al. (2015) also proposed a coherence measure that outperformed the existing ones, called C_V. The C_V metric is used to evaluate the topic models in this paper.

Other common ways of evaluating topic models is through human interpretation and visualisation. The topics need to be understandable from a human perspective too. Visualisation can help present other aspects of the topics, such as similarities. Sievert's (2014) work has become a standard practice to visualise the topics onto a two-dimensional plane.

3 Methodology

3.1 Data

For this research, the data is personally collected using the real-time Twitter API. 89,000 tweets were extracted, which corresponds to 17.5 MB of data. Twitter provides access to a variety of data variables, but this research will only extract seven. Those are referred to in Table 1.

1	Tweet_ID
2	Tweet
3	Created_at
4	Coordinates
5	Language
6	Source
7	User_ID

Table 1: Extracted features

The tweet ID serves as the primary key and is unique to the post. Furthermore, the research requires the longitude and latitude of the Tweets posting location. This is only visible if the user has their Twitter location turned on, otherwise it will return a *null*. Regardless if the tweets return coordinates, the content of all tweets will be used for the modelling of the topics, as more data often leads to better results. However, for the geographical zoning of the interest, only the geotagged tweets can be used. It would be possible to dummy code the coordinates for tweets that have those missing. For example, there are ways to return coordinates with a given location (Carriere; n.d.). However, that approach would not be practical as this research only focuses on one city. Thus, many observations would return as *'Dublin City, Ireland'* and this would not be granular enough for this research.

The language of the tweets is also extracted. The aim of this research is only considering the English language; therefore, non-English tweets can be filtered out by utilising this variable. This will become more clear in section 3.4.

To identify the areas, an additional dataset is required. This is a shapefile of ungeneralised small areas for Ireland. [4]. Small areas are the most granular geographical level of Ireland, where each small area consists of 80 to 120 dwellings (CSO; 2016). Each tweet's coordinates relate to a specific small area for which the topic zones can be identified. More information is provided in Section 4.For figures 5 and 6, the generalised version of the shape file is used because this is a more compact version for mapping, as stated by the CSO (2016). [5]

3.2 Data Collection and Extraction

The data is collected using the public Twitter streaming API. Twitter developed and maintains this API for several use cases, one of which is to evaluate Twitter data to

[4]Small Areas Ungeneralised: `https://data-osi.opendata.arcgis.com/datasets/c85e610da1464178a2cd84a88020c8e2_3`

[5]Small Areas Generalised:`https://data-osi.opendata.arcgis.com/datasets/68b14cef8cf247b191ee2737e7e6993d_1`

inform business decisions. This will be the use case for this analytical research. In order to interact with this API, one must fill in a form on developer.twitter.com to obtain API credentials. For this data collection, the Python library called Tweepy is used to interact with the streaming Twitter API. [6]

A Python script is written to specify the extracted variables and to store the data in a temporary SQLite database. To specify tweets that were posted in Dublin County, the script filters by a specific bounding box. All tweets within this bounding box are collected. However, only tweets with location tracking on will return coordinates. The script is kept running for the entirety of the data collection period of two weeks. This database is not used at any other stage but serves the sole purpose of safeguarding the data in case of, for example, a power cut. Databases are designed to commit the inserted data and hence, are more reliant to store the data than text files. Nonetheless, a copy of the database is made and exported into a csv file, after which it is uploaded onto the cloud as a backup.

3.3 Data Transformation

After the data is collected and stored, the data is transformed. This step was initially not considered, but after assessment from the Ethics Committee, necessary steps were taken to anonymise the tweet ID's and user IDs by hashing the ID's to randomised ID's. This makes it so that ID's are not identifiable anymore. The transformation is performed using a *sha3_224* hashing algorithm, which converts the ID's into a random string of utf-8 characters. Furthermore, the age of a user cannot be identified, as Twitter does not release this via any of their API's. Regardless, Twitter does not require all users to submit their age (Pearson; 2018).

3.4 Data Pre-Processing

Before being able to use the topic modelling algorithm, it is important to build a corpus of text whereby the text has been pre-processed. The first step that is taken is ensuring that all text that will be processed are written in English. With the Twitter API, the language of the tweets is retrieved. Therefore, the non-English tweets can be disposed. After this step, 74,599 tweets remain.

Tweets can be retweeted posts from other users. In this case, the retweeted post will have the letters RT in it. Those are removed. Also, the removal of mentions of Twitter handles. Those are usually in the replies to other tweets. Tweets can also have URL links embedded in them. In these cases, the links are removed. Characters like apostrophes and hash marks are also removed. Hashtags are words within a tweet that users want to stand out. The hash marks are removed, but the actual words are not because they may provide necessary information on the topic of the tweet. All non-utf8 characters, such as emojis cannot be used in the topic modelling algorithm and hence, are also removed. Furthermore, numbers are taken out, because they do not serve any real meaning to the content of a tweet. Unfortunately, this may remove cases like *'gr8'* or *'some1'*. Lastly, stop words are removed and the remaining words are lemmatised. An additional step was undertaken in an effort to improve the model afterwards; profanity is also removed from the text.

[6] Tweepy documentation: `http://tweepy.readthedocs.io/en/v3.6.0/`

After all the cleaning actions, the words are tokenised and a bag of words dictionary is created. As an extra measure, words in the dictionary must appear in at least 15 tweets and no more than half the size. From this dictionary, a corpus matrix is created, which converts the words into integers by the location and frequency across all tweets. At this stage, the data is ready to be used for topic modelling.

3.5 Evaluation Methodology

The methodology of evaluating the results will follow similar approaches as references in the literature review. The C_V metric for coherence is used one way that the topics were evaluated. The general rule of thumb is higher is better. But since there is no way of saying what is good or bad, you must take the value at face value. In other words, it is best to compare with other models to improve the model. However, caution should be taken when stating that a model is good or bad based on the measure value alone. The Gensim library has implemented most of Roder's research (Rehurek and Sojka; 2010)(Roder et al.; 2015).

Human interpretation is needed to validate the understandability of the topics and their respective words. As context is important in language, human interpretation is a significant factor. In this paper, human interpretation is assessed using a Likert scale by the author. This may introduce a certain extent of subjectivity. The topics are also evaluated with the use of pyLDAvis (Mabey; 2015), which is a tool based on Sievert's (2014) work. It visualises the topics to inspect their similarities; see figure 4.

Furthermore, the approach is evaluated by cross validating the model five times, which is consistent with Anagnostopoulos et al. (2017). "We evaluate our approach performing a 5-fold cross-validation" (Anagnostopoulos et al.; 2017, p.529). Lastly, to test the model, new, unseen data is inserted. The model gets updated afterwards with the new data.

4 Design & Implementation

The design of the implementation will function hierarchically. Figure 1 provides a simplistic overview of this. The input data has been pre-processed and transformed at this stage. From this, the text is used to divide the tweets into topics. This process is performed using the LDA algorithm. Like the pre-processing, the modelling was scripted in Python with the use of the Gensim library (Rehurek and Sojka; 2010). This library has many integrated features for topic modelling, including the LDA algorithm. The modelling phase consists of tweaking the parameters of the algorithm on the derived dictionary and corpus from the data. Once the model is built, the topics are labelled, and they can be linked back to the documents. A drawback is that labelling can be subjective. A Likert score aims to address this to an extent.

Each document is assigned the topic that has the highest probability. Afterwards, the geotagged tweets are used to create zones within Dublin. There are not as many geotagged tweets, which is further reduced by quality issues of certain coordinates, leaving 2,722 of the 74,599 tweets to be zoned. A geometry object is created from the shapefile of ungeneralised small areas that is then used to assign a small area to each tweet. When a small area has more than one topic, the average of the probabilities of the topics is calculated. The topic with the highest average is then assigned to the small area. In the end, 860 unique areas were identified. This concludes the last step of the implementation.

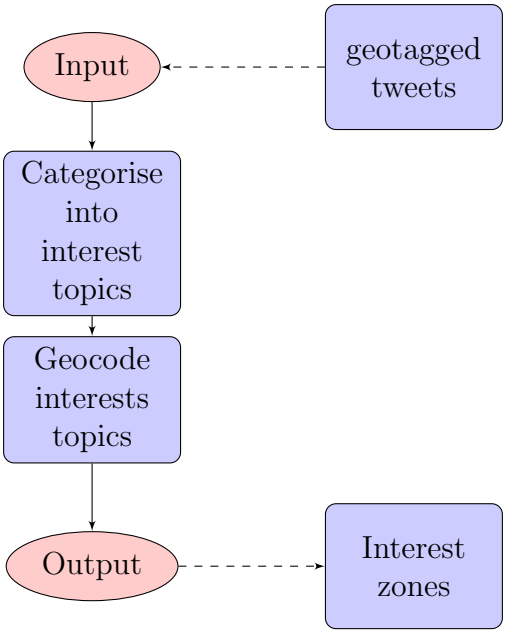

Figure 1: Design of Implementation

Lastly, a careful design specification was considered in collecting the data. Due to the research period being constrained, the collection could take place over a maximum of two months. However, those months saw several significant events in Ireland and therefore, the actual data collection cycle was determined to avoid those events. The collection period began after the Irish referendum of 2018 and ended before the start of the World Cup. The collection period was carefully determined to reduce the frequency of one-sided topics, such as politics and sports.

4.1 Estimating k-topics

This section refers to the design implementation that was chosen before the topic modelling phase. One of the parameters of the algorithm that is required to be set is the number of topics (k) that will be returned. In other words, the algorithm determines the topics based on the number of topics that is required to return. Like many other machine learning approaches, the value for k will influence the results of the model. Therefore, it is of significance to approximate the ideal k-topics.

The C_V coherence score will help estimate k. This score has a range 0 to 1. In order to be impartial, a default algorithm was run, only changing the number of topics, and then calculating the scores. As can be seen in figure 2, the coherence score gradually increases when k gets larger. However, after 20 topics, the coherence score is no longer reliable.

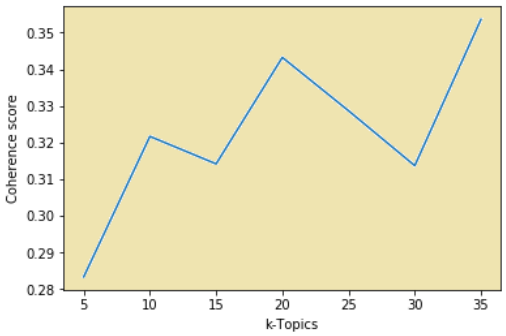

Figure 2: Estimating k using coherence metric

In practice, the model was first trained with $k = 20$. The chunk size parameter is used to specify the number of documents that the algorithm will parse through each time. For this research, it was set to 7500 documents because this is approximately 10% of the overall corpus. In the algorithm, the number of passes defines the number of times that the algorithm will run through the same chunk. The number of passes is changed accordingly in order to improve the model. Throughout the modelling phase, the number of topics was changed. The model was trained on 10 and 15 topics as well, because they would not decrease the coherence scores that much, in accordance with figure 2, whilst also keeping mindful that a lower number of topics will improve the finding interpretable topics. A large number of k may also result in too many similar topics or sub-topics. Hence, it was decided that models with 10, 15 and 20 topics were trained.

5 Evaluation & Results Analysis

5.1 Results

As can be seen in figure 3, the variable *topics10* is significantly higher for all passes than its counterparts, bearing in mind that the coherence score may differ for each run. Not only did the models with 10 topics perform better in terms of coherence scores, the top words for those models were generally more interpretable and semantically fitting. Furthermore, increasing the number of passes did not significantly improve the coherence scores, nor did they improve the interpretability of the topics. In fact, certain models were trained with passes 10 times higher and they did not have a significant impact.

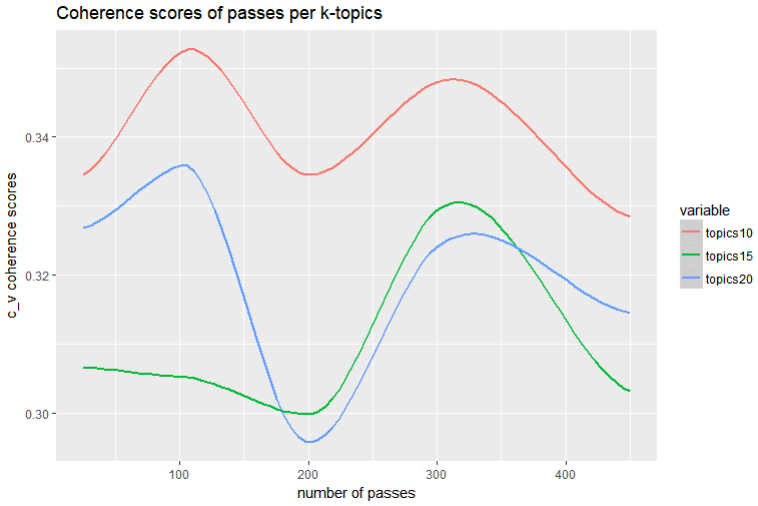

Figure 3: Estimating k using coherence metric

The final model is also a model with 10 topics, however, it was determined not only based on the coherence scores, but also via the human interpretability and visualizing the topics. The final model was trained with 50 passes. The parameters for alpha and eta are on default, as they did not improve the model.. The model failed to produce a high coherence score, but it had interpretable topics of which were diverse in nature too. The results of this model are shown in table 2. In no specific order, the topic number represents the same number in figure 4. Additionally, a Likert score has been assigned to each topic based on their human interpretability, where 1 represents incoherent/uninterpretable words and 5 represents topics that have clear representative words.

Results of Final Model			
Topic #	Label	Top 10 representative words	Score
1	Undefined	0.025*"get" + 0.015*"one" + 0.015*"see" + 0.014*"day" + 0.013*"time" + 0.013*"think" + 0.012*"make" + 0.012*"good" + 0.011*"know" + 0.011*"people"	1
2	Event	0.055*"ireland" + 0.041*"world" + 0.026*"irish" + 0.026*"team" + 0.022*"cup" + 0.020*"meet" + 0.017*"morning" + 0.013*"beautiful" + 0.013*"money-conf" + 0.012*"event"	4
3	Job	0.111*"dublin" + 0.078*"great" + 0.075*"look" + 0.037*"job" + 0.025*"anyone" + 0.020*"county" + 0.018*"enjoy" + 0.018*"interest" + 0.017*"forward" + 0.017*"check"	4
4	Sports	0.044*"best" + 0.035*"please" + 0.034*"play" + 0.033*"next" + 0.032*"win" + 0.031*"game" + 0.025*"tomorrow" + 0.024*"yes" + 0.020*"tonight" + 0.019*"ticket"	5
5	Politics	0.094*"love" + 0.021*"via" + 0.017*"brexit" + 0.017*"pay" + 0.015*"vote" + 0.015*"break" + 0.014*"believe" + 0.011*"party" + 0.011*"stand" + 0.011*"lead"	5
6	Livestyle	0.102*"thank" + 0.057*"new" + 0.032*"happy" + 0.032*"wait" + 0.025*"support" + 0.021*"loveisland" + 0.017*"excite" + 0.016*"weekend" + 0.015*"news" + 0.015*"brilliant"	2
7	Home	0.036*"house" + 0.029*"bed" + 0.017*"service" + 0.015*"agree" + 0.014*"road" + 0.014*"child" + 0.013*"care" + 0.012*"song" + 0.012*"build" + 0.011*"state"	4
8	Nightlife	0.076*"year" + 0.045*"last" + 0.045*"week" + 0.043*"live" + 0.037*"night" + 0.019*"lose" + 0.017*"hour" + 0.015*"link" + 0.014*"drink" + 0.014*"since"	3
9	Social	0.053*"watch" + 0.030*"miss" + 0.029*"join" + 0.017*"soon" + 0.016*"video" + 0.014*"mate" + 0.013*"medium" + 0.012*"market" + 0.011*"social" + 0.011*"pic"	3
10	Job	0.051*"open" + 0.045*"amaze" + 0.030*"trend" + 0.029*"latest" + 0.021*"click" + 0.021*"view" + 0.021*"true" + 0.020*"hire" + 0.019*"unite" + 0.017*"apply"	3

Table 2: Model with 10 topics and 50 passes

The probabilities assigned to each word represent the contribution toward the topic. It can be interpreted as the weight that a word gives to the meaning of the overall topic. Topic 4 is labelled Sports because the combined connotation of the words insinuates

sports. However, it was difficult to label topic 1, because the words have no combined interpretation. Several topics have words that can fit any topic, such as *brilliant* in lifestyle.

Figure 4, generated with pyLDAvis, represents the model on a distance map. The visualisation serves as a way to judge the models on their similarities. It is ideal to have relatively equal topics in size and to have the topics spread out evenly, hence indicating that the model can create distinct topics.

Figure 4: Distance Map of Similarity

Undefined [1] is much larger with respect to the other topics. This may be because it entails many words from documents that do not have a defined coherent topic, and those words are more frequent in the corpus. Also, *lifestyle* [6] and *nightlife* [8] are overlapping, which makes sense because the words are similar in nature. Topics 3 and topics 10 are the only two topics that have a similar label, namely job-related. On the distance map, it is also evident that those topics are closely related.

Figures 5 and 6 is the result of the interest zoning. To derive the topics from the tweets, the model did not utilise the coordinates as a feature, meaning the model did not know which documents had coordinates. Nonetheless, topic 1 has predominately more tweets with coordinates turned on. In fact, most small areas in which there was tweeted from have topic 1. As seen in figure 4, this could be explained by the large proportion of

tweets assigned to topic 1, regardless of whether they have coordinates. Therefore, it is more likely for geotagged tweets to be in topic 1 than the same for other topics.

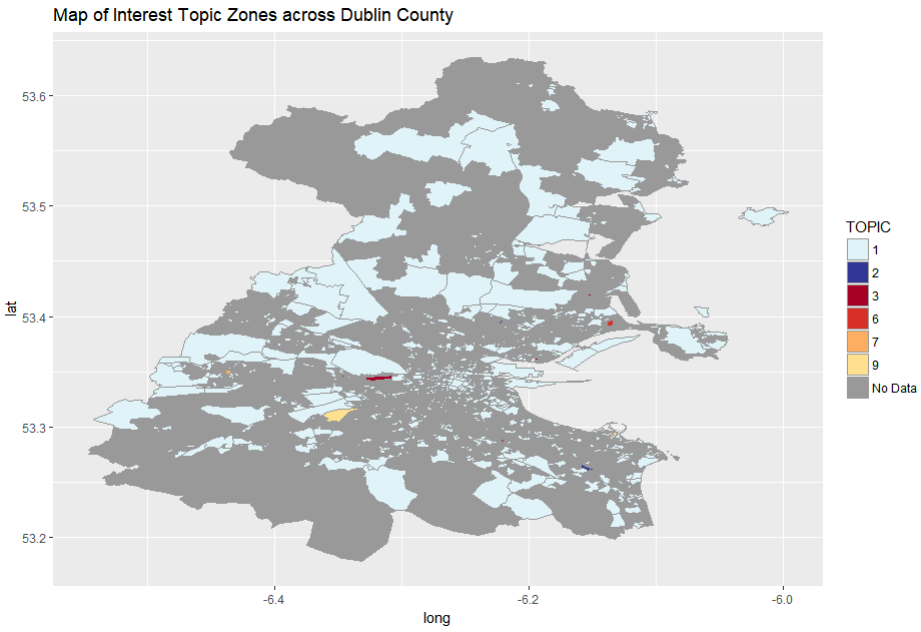

Figure 5: Map of interest zones across Dublin County

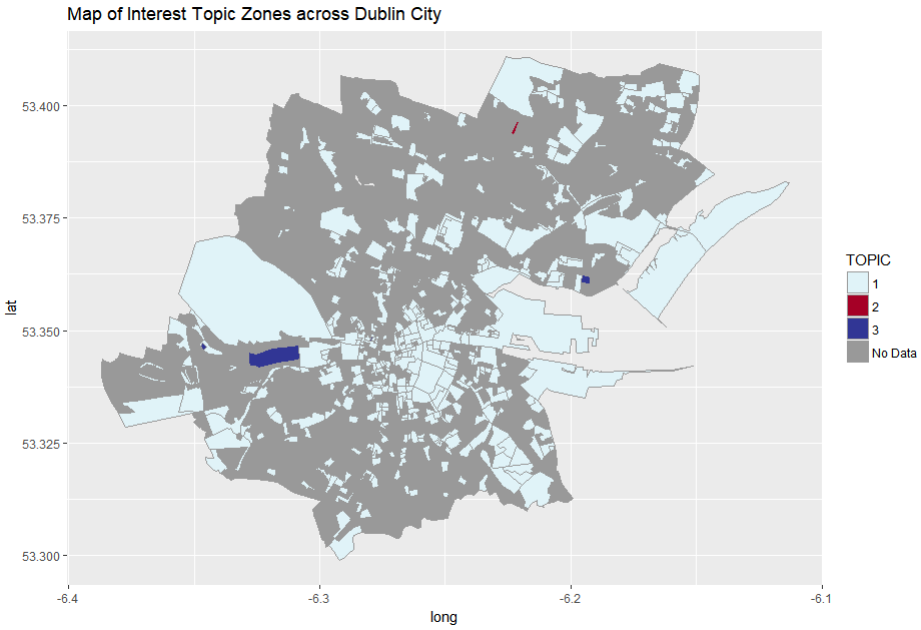

Figure 6: Map of interest zones across Dublin City

The lack of geotagged tweets is clearly visible in the figures. Due to this, there is no fair representation of the topics across Dublin. While true, the framework for zoning would serve its purpose better if refined topics had more geotagged tweets. Thus, the mediocre zoning can partially be explained by the small number of geotagged data.

5.2 Insights

As Twitter is an online medium, it is used to converse social themes that could be discussed in a face-to-face conversation as well. Twitter is a theme-based medium with common re-occurring themes and thus, those themes are reflected in the topics. From running numerous models with different sets of parameters, it was evident that common topics came back. Those were topics about UK politics, job related, sports, but also many topics about being excited for an event, like a birthday or the weekend. Furthermore, Twitter is also frequently used to run advertisements. For example, a common re-occurrence is the job application advertisements. Such advertisements may not be a fair representation of what Dubliners talk about. The author chose to keep those tweets in the corpus, because otherwise it would add additional complexity to remove such occurrences as it may result in the removal of important latent patterns of other tweets.

See our latest #Dublin, County Dublin #job and click to apply: Test Engineer - https://t.co/IgZZjr26k3 #QA #Hiring #CareerArc

Table 3: Example of job advertisement tweet

Another finding is that the algorithm would frequently pick up words from completely opposing documents and combine them under one topic. The most captivating example of this is a topic produced which included the following top words:

world + cup + via + worldcup + vote + trump + russia + deserve

Table 4: References to football and politics

At first, it may appear to be incoherent, but there is an explanatory link. The algorithm picked up latent patterns between documents about the world cup in Russia and documents regarding the Trump-Russia investigation. This proves how latent patterns are defined by the algorithm.

In terms of model specifications, there were also several findings. Increasing the number of passes, or the number of times the algorithm gets to see the documents, does not necessarily improve the coherence scores or the human interpretability of the topics. Furthermore, this is a trend that was also noticeable for a lower number of passes. For example, passing it 10 times gave better coherence scores, but the top words per topic were not interpretable, so no real topic label could be assigned and therefore, left undefined. In fact, models with lower coherence scores produced better distinguishable and interpretable top words for the topics, thus producing better semantically *correct* topics. Furthermore, lowering the number of topics from 20 improved the coherence scores of the models. With 10 topics, it appeared that it is easier to divide the documents into distinct topics. With a smaller number of topics, it is also evident that there are fewer topics with the pointless representative words. In other words, it is easier to label a higher proportion of the topics when the algorithm is set for fewer topics, whilst the coherence scores are also generally higher. Hence, it makes it easier to create distinct topics. With a higher number of topics, there are more repetitive topics, for example several topics about sports. Moreover, with a higher number of topics, there are more undefined topics too. Overall, the coherence scores are higher for the 10 topic models as opposed to the 15 to 20 topics in the models. So, from both the metric and human

interpretability point of view, fewer topics produced more coherent topics for this tweet analysis.

5.3 Discussions

Firstly, it is evident from the results that increasing the number of passes provides little improvement. Increasing the number of passes increases the computational time, because the algorithm iterates through the data more. The improvement and computational time trade-off informs that training the model for longer and more complex times, did not give significant benefits. In fact, the final model trained on only 50 passes and performed just as well, but with only a fraction of the training time.

While the model was able to detect certain topics for many of the tweets, it is questionable whether all tweets really do belong to a specific topic. The coherence scores for all models are in and around 0.30-0.34. Even though topics could be defined from human interpretability, there are still questionable words in the top words for certain topics. This indicates that the topics are still relatively weak. Changing the passes and other parameters did not significantly improve the model. This may indicate that topic modelling is not adequate with the text of tweets.

Tweets are difficult to assign a single topic to, as there can easily be several topics per tweet, but even more importantly there are many sub-topics within each topic. For example, a tweet about a football player may not necessarily be about sports but could be a personal tweet. However, assigning the sports label to this particular topic could also put other tweets from the same topic under this sports label, even when it is unrelated to sports. This seems to be the biggest factor influencing the model's results, as the algorithm sees relationships between documents that may share similar words but not the meaning, which is an issue to most NLP algorithms. The quality of the topic is dependent on the quality of the data. There is a dilemma: either be conservative and risk losing patterns or be liberal and risk introducing more noise. Perhaps, an additional cleaning step can be done. Correcting spelling mistakes may improve the quality of the text (Garbe; 2018).

Realistically, the theme or topic of a piece of text will be easier to determine when there is more text because it allows for important words to occur more frequently. Moreover, more words could increase the relationship between words within the same document and across others.

It is also evident that certain topics do not serve a real purpose with regards to identifying topics for advertising, for example *Social*. Twitter is often used to express thoughts about everyday life without giving the context, or it can simply have conversations between acquaintances on a topic that would be considered irrelevant to an outsider. Having a way of identifying such tweets and removing them may reinforce the more relevant topics. Furthermore, this would also help with the previous argument, as the algorithm can detect more distinct topics. For advertising purposes, it is unlikely that this specific model would be able to function. However, it might provide insight in certain areas in the city.

The *Lists* approach by Anagnostopoulos et al. (2017) will have static topic names, whereas the topic modelling approach will have dynamic topic names. This is because the subjects being tweeted about will be event based. Text from tweets are very event based. In other words, the data in this research included many tweets referencing Brexit and World Cup, hence, top words for topics will have words about those events. If the same

topic modelling was done on tweets for next year, it will most likely reference different events. Therefore, the topics' names will be dynamic. This could lead to issues, as the specific areas could change topics based on the event. However, the *List* approach is more stable when identifying topics in a city's zones. The *List* approach also provided better distinct zones, whereas the topics modelling approach mapped mainly topic 1. This is due to the lack of distinct and clear topics from this approach and the small number of geotagged tweets. The *List* approach offered more variety in terms of the topics, and this is represented in the difference of this research's and Anagnostopoulos et al. (2017)'s maps. However, Anagnostopoulos et al. (2017) had access to more geotagged tweets due to Twitter Firehose.

The small areas were chosen as the granular level because these are the same areas for the Census data. This means that advertisers could potentially filter, not only by topics, but also on a socio-demographic level. Additional geographical data may also provide extra information, such Dublin Dashboard.

6 Conclusion and Future Work

Topic modelling using tweets is not ideal, because there are many topics and sub-topics that can occur within tweets. For example, you can have many different types of conversations between users. There are also many advertisements of companies on Twitter, which make it difficult to get distinct topics. It is complicated to create interpretable and cohesive topics, even more so when dealing with many documents with little text. Furthermore, it is also complicated to create a handful of distinct topics with many different styles of text. In other words, it is difficult to classify the text because there are many different scenarios of text, and therefore, there are too many themes than possible for the purpose of this research.

Another conclusion is that there are not many users that have their coordinates turned on. This results in an unfair representation of the population in terms of tweets, and therefore topics. The small number of geotagged tweets hinder the effort to zone the interest topics well.

Nonetheless, the topic modelling of tweets for interest zoning has potential, but it requires optimisation. Two solutions are proposed: first, a way to identify and remove tweets from the dataset that are unusable for topics should be further explored. In other words, those that cause noise in the modelling, like advertisements. A second solution would be to have more geotagged tweets. A larger dataset would not be improper, but it may not solve the problem either. Ideally, a way to estimate coordinates, like Cheng et al. (2010), but with shorter range, would be favourable.

The *Lists* approach. may be superior, as it is more refined, trustworthy and has static interest categories, but it has its limitations as well, some of which the topic modelling addresses. The two approaches complement each other. Future work may explore an ensemble technique of the two approaches, using the efficiency of the topic modelling and considering the actual text, but keeping in mind the user profiling of the *Lists* approach. A combination of both techniques could be beneficial for more accuracy in future research. Future work may also address a way to identify tweets that were posted whilst commuting. Tweets during commute may not represent the true interest of the citizens living in that particular area.

Lastly, future work may be explored with other media than Twitter, for instance Face-

book or Instagram. Those platforms, however, would introduce the additional obstacle of images because text is not the prime focus, unlike tweets . Perhaps, a very advanced proposal could incorporate the textual data with image recognition and processing.

References

Anagnostopoulos, A., Petroni, F. and Sorella, M. (2017). Targeted Interest-Driven Advertising in Cities Using Twitter., *ICWSM* pp. 527–530.

Annual Report (2018). *Technical report*, Twitter, Inc., San Francisco, California.
URL: *https://investor.twitterinc.com/annuals-proxies.cfm*

Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003). Latent Dirichlet Allocation, *J. Mach. Learn. Res.* **3**: 993–1022.
URL: *http://dl.acm.org/citation.cfm?id=944919.944937*

Bruner, G. and Kumar, A. (2007). Attitude toward Location-based Advertising, *Journal of Interactive Advertising* **7**.

Carriere, D. (n.d.). geocoder documentation.
URL: *https://geocoder.readthedocs.io/*

Chang, J., Gerrish, S., Wang, C., Boyd-graber, J. L. and Blei, D. M. (2009). Reading Tea Leaves: How Humans Interpret Topic Models, *in* Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams and A. Culotta (eds), *Advances in Neural Information Processing Systems 22*, Curran Associates, Inc., pp. 288–296.
URL: *http://papers.nips.cc/paper/3700-reading-tea-leaves-how-humans-interpret-topic-models.pdf*

Chen, J., Nairn, R., Nelson, L., Bernstein, M. and Chi, E. (2010). Short and Tweet: Experiments on Recommending Content from Information Streams, *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, ACM, New York, NY, USA, pp. 1185–1194.
URL: *http://doi.acm.org/10.1145/1753326.1753503*

Cheng, Z., Caverlee, J., Barthwal, H. and Bachani, V. (2014). Who is the barbecue king of texas?: a geo-spatial approach to finding local experts on twitter, ACM Press, pp. 335–344.
URL: *http://dl.acm.org/citation.cfm?doid=2600428.2609580*

Cheng, Z., Caverlee, J. and Lee, K. (2010). You are where you tweet: a content-based approach to geo-locating twitter users, *Proceedings of the 19th ACM international conference on Information and knowledge management*, ACM, pp. 759–768.

CSO (2016). Census 2016 Boundary Files - CSO - Central Statistics Office.
URL: *https://www.cso.ie/en/census/census2016reports/census2016boundaryfiles/*

Garbe, W. (2018). SymSpell: 1 million times faster through Symmetric Delete spelling correction algorithm. MIT License. original-date: 2014-03-25.
URL: *https://github.com/wolfgarbe/SymSpell*

Hoffman, M. D. and Blei, D. M. (2010). Online Learning for Latent Dirichlet Allocation, p. 9.

Liu, L., Tang, L., Dong, W., Yao, S. and Zhou, W. (2016). An overview of topic modeling and its current applications in bioinformatics, *SpringerPlus* **5**(1).
URL: *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5028368/*

Mabey, B. (2015). pyLDAvis documentation.
URL: *http://pyldavis.readthedocs.io/en/latest/index.html*

Manning, C. D., Raghavan, P. and Schtze, H. (2008). *Introduction to Information Retrieval*, Cambridge University Press, New York, NY, USA.

Nugroho, R., Yang, J., Zhao, W., Paris, C. and Nepal, S. (2015). What and With Whom? Identifying Topics in Twitter Through Both Interactions and Text, *IEEE Transactions on Services Computing* pp. 1–1.
URL: *http://ieeexplore.ieee.org/document/7906613/*

Pearson, J. (2018). Twitter Is Banning Anyone Whose Date of Birth Says They Joined Before They Were 13, *Motherboard* .
URL: *https://motherboard.vice.com/en_us/article/vbq3dm/twitter-banning-anyone-under-13-date-of-birth-gdpr*

Rehurek, R. and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora, *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, ELRA, Valletta, Malta, pp. 45–50.

Roder, M., Both, A. and Hinneburg, A. (2015). Exploring the Space of Topic Coherence Measures, *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM '15, ACM, New York, NY, USA, pp. 399–408.
URL: *http://doi.acm.org/10.1145/2684822.2685324*

Sievert, C. (2014). LDAvis : A method for visualizing and interpreting topics, *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, Association for Computational Linguistics, Baltimore, Maryland, USA, pp. 63–70.
URL: *http://www.aclweb.org/anthology/W14-3110*

Sorella, M. (2018). Targeted Interest-Driven Advertising in Cities Using Twitter [email].