Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

School of Computer Science and Statistics

Validating an infrastructure for online chatter monitoring

Deepanshu Jain

August 19, 2022

A dissertation submitted in partial fulfilment
of the requirements for the degree of
M.Sc (Computer Science : Data Science)

Declaration

I hereby declare that this dissertation is entirely my own work and that it has not been submitted as an exercise for a degree at this or any other university.

I have read and I understand the plagiarism provisions in the General Regulations of the University Calendar for the current year, found at `http://www.tcd.ie/calendar`.

I have also completed the Online Tutorial on avoiding plagiarism 'Ready Steady Write', located at `http://tcd-ie.libguides.com/plagiarism/ready-steady-write`.

Signed: _____ Date: 19/08/2022

Abstract

In today's day and age, the number of people in the third stage of their life is increasing, and so are the expenses required to look after them and care for them. This increase has called for massive growth in the demand for technology to assist the elderly or help others to take care of them. But regardless of the money being invested and the efforts made, an essential factor that needs to be determined is how well the older generation would adapt to these technologies. Support groups have dominated all the research surrounding the interaction of older people and technology. But we live in an age of social media. COVID-19 forced more senior people to come out of their shells and get their hands dirty with technology since technology remained the sole communication medium during the social lockdown. And with the development of advanced text analytics techniques, it will be interesting to see whether the results found by the previous researchers in this field are still valid. And since society plays a significant factor in influencing the elderly to use technology, we will also look at the emotions of people talking about older people.

We created a pipeline that can be used by anyone trying to fetch data from specific social media platforms, which gave an accuracy of around 80% on three different platforms. We then used Part of Speech Tagging techniques to find out what parts of speech were used the most by either older people or people talking about older people. We saw increased use of Possessive Pronouns and compared the same with data extracted using keywords not related to older people to confirm the same. We then tried to mine the emotion behind the texts and found that for nearly half of the keywords used across the three social media websites (17 out of 34), the primary feeling shown was that of being scared. The second most widely exhibited feeling across the dataset was the feeling of joy when the keywords were related to welfare schemes of the government for older people. For a few keywords, the emotion was sadness and angriness. Our research showed contradicting results to the outcomes delivered by previous researchers who used focus groups to get an idea of the emotions of older people towards technology.

Acknowledgements

I cannot comprehend the amount of learning I have done in the past year, especially during this research dissertation. I will remember the time I have spent here, not just for my near future but my whole life. I want to take this opportunity to express my immense gratitude to the few people without whom this dissertation and the journey of my master's degree would not have been possible.

First and foremost, I would like to show immense gratitude toward **Dr. Carl Vogel** for his constant support, sincere and selfless advice, and prompt assistance, without which this research would not have been possible. I would also like to thank **Dr. Erwan Moreau** for his guidance in the latter stages of the study.

In addition, I would also like to thank my friends and housemates **Archit Jhingan**, **Akul Rastogi**, **Angkirat Singh Sandhu**, and **Sakshi Boolchandani** for tolerating me during the hard times and making sure I was never too stressed with work. I also want to thank **all my friends** who made sure that I never missed my family and made me feel like having a family away from home. I also owe a sense of gratitude to my brother from another mother, **Sagar Chugh**, who has always been there for me no matter what.

Last but certainly not least, words are not enough when it comes to describing the appreciation I have towards my parents (**Mr. Minaksh Jain** and **Mrs. Charanjeet Jain**), my sisters (**Jaspreet Jain Anand** and **Kalyani Jain Babbar**), my brothers-in-law (**Kartik Anand** and **Ankit Babbar**) and the newest member of my family, my nephew **Grahill Babbar**. Their immense love, unconditional support, and tolerance made this journey worthwhile. Without them, I don't think I would have been able to handle the last year and overcome the difficulties I faced.

Contents

List of Figures

List of Tables

1 Introduction

This chapter 1 will deal with introducing the topic of the thesis as well as the motivation behind choosing the same. We will try to reason why the research being undertaken is necessary and what all will we will doing to take this research forward.

1.1 Motivation

Internet usage in the world has increased drastically in the past few years, with the total global users reaching 4.95 billion by 2022 or 62.5 percent of the worldwide population, per the Digital 2022 Global Overview report [15]. This trend can also be associated with the fact that the internet became one of the significant ways of communication with one's family during the pandemic. According to the reports by the Central Statistics Office of Ireland, as of 2021, nearly 93% of households in Ireland had internet connections. However, the fact remains that the majority of people that use the internet are between the ages of 16 and 45. On the contrary, nearly half (46%) of people above the age of 75 have never used the internet [3]. Internet, for them, is part of a more severe and significant problem: their fear and overall attitude towards technology. In their study, Rachel et al. [22] also found that, despite 77% of adult Americans owning a handheld device, one-third of the older population had never used the internet; only 42% owned a smartphone, and just 35% knew how to use a tablet.

Furthermore, the increase in the costs of providing care for older people and people with disabilities combined with the rapid pace at which the technologies have advanced has led to a massive boost in the research for technologies that could assist with the same [22]. Nevertheless, an essential factor that remains unknown is whether the technologies being developed would be accepted by the people or not. There must be a partnership between the people developing such technology and those from whom that technology is being made to understand what makes them skeptical towards technology. The success of the said research would depend on the acceptance of the older people and the people who will have to interact with those technologies.

Several age-related factors impact the acceptance of technology by older people, including

cognitive, social, physical, financial, emotional, and educational. These factors not only influence how older people live their lives and the quality of their lives but also affect how they use technology [22]. However, people's reaction to aging is very subjective, and access to care that can support them with the changes is not always possible. Especially for people who are either socially isolated for many reasons or cannot afford such services in general. Nearly 80% of the people desire to live in their own homes and within their communities rather than spending their lives at an assisted care house [8]. Hence, the technologies coming up to assist older people play a significant role in how they can assist them and how they can be made more accessible to older people.

To better understand the acceptability of computers and technology in general, significant work was done by Davis et al. [7], introducing the Technology Acceptance Model (TAM), which was based on the Theory of Reasoned Action (TRA). TRA suggests that an individual's actual behaviours are motivated by the behavioural intention to perform that behaviour. This intention is, in turn, influenced by the attitude towards behaviour and norms that subjective individuals hold themselves to, which can be considered as the perceived social pressure to perform or not perform that task. Davis took this theory and applied it to technology, stating that these intentions are affected by the perceived usefulness of the technology and one's attitude towards technology in general. They also introduced a parameter called precise ease of use which measured the degree to which a technology is free from the effort to a target user. Currently, in its second iteration, the model suggests that seven main factors affect how an individual uses technology: Price value, Habit, Social Influence, Effort Expectancy, Hedonic Motivation, Facilitating Conditions and Performance.

Based on the work of David et al. on the general technology acceptance model, multiple steps have been taken to understand the acceptance of older people towards technology. Out of all the models, the major ones are STAM or Senior Technology Acceptance Model by Chen Chan [4] and STAAM or Senior Technology Acceptance and Adaptation Model by Renaud et al [23].

In STAM, it was postulated that the acceptance of technology in older people depends on personal characteristics (i.e., gender, social status, age, economic status), level of anxiety, self-efficacy, and other facilitating conditions such as cognitive ability and health. A critical finding of STAM was that acceptance of the technology was not affected by the perceived usefulness and ease of use. STAAM, on the other hand, suggested that acceptance and adoption came in three phases: Objectification, Incorporation and Acceptance. Objectification is when the intention of using the technology is defined; it is influenced by social factors and the usefulness of the technology. Incorporation where the older people incorporate the technology into their daily lives to understand whether it is functional. Acceptance or Rejection depends on the attitudes formulated during the first two steps.

No matter how evident the above-stated factors might appear to be, it cannot be questioned that technological acceptance is an intricate subject, especially when we talk about older adults. Hence, knowing what factors facilitate or hinder acceptance is not enough. It is also crucial to know people's general attitudes towards such technologies. It is also essential to know what other people think about older people using technology since society plays a meaningful role in accepting technology at any level.

Furthermore, despite the generational technological divide showcased above and how the older generation uses a handful of technological devices, it is rather strange how so many researchers have shown that older people have opinions that are utterly contrasting to stereotypes of them being scared and not having a positive outlook towards technology [22] [10] [30] [8].

This discrepancy might be possible because of how the experiments for the literature were set up; most of the research was done using Focus groups, usually consisting of 15-20 people. The issue with focus groups is that since the people usually come from similar demographics, the results can be very skewed. Moreover, they can also be heavily inclined towards one side of the discussion, and the reasons for that will be discussed later. Hence, it becomes vital that when we try to get people's opinions towards older people using technology or the idea of older people towards technology, we do not do it using focus groups. Social networking platforms jump out as the best option where we can get people's opinions on almost every topic; hence, we will use the same.

1.2 Research Question

The main idea of this research is to check the opinion of older people towards technology and the opinion of other people towards older people using technology since social factors play a meaningful role in the adoption of technology amongst older adults [10] [30] [12]. However, the data will have to be extracted from various media platforms before that can be done. For this, a reusable data pipeline will be created that can later be used by anyone with little or no coding experience to extract data from the social media platforms we will be working with.

We will look at various ways data can be extracted from the platforms and select the most appropriate one for each platform. We will pick several keywords and extract data from the selected platforms based on those keywords. The data then will be parsed into a structured format such as CSV or Spreadsheets, cleaned and pre-processed. Pre-processing includes several steps like removing duplicates, deleting stop words, eliminating unnecessary information, getting rid of punctuations, stemming / lemmatizing etc.

Data organizing and cleaning play a considerable role in any project, and according to studies, nearly 60% of the time is spent organizing the data. Nearly 19% of the time is spent collecting

it, meaning more than 75% is spent just collecting and refining the data [21]. The pipeline that will be created can reduce that time drastically if someone has to extract data from the chosen sources. Hence, they will be able to spend more time designing the models, analysing the data, and refining the models to get the best results.

The cleaned data will then be sent to a Part of Speech Tagger, which is used to provide tags to each word in a text. It is done as a pre-processing step in the NLP pipelines to provide context to the algorithms so that they understand the text better. We would review different POS taggers and select the one best suit our data. Then a principal component analysis (PCA) of the data will be done. PCA generally reduces high dimensional data into low dimensional data with the same or equivalent variance. We will try and find the words used most throughout the texts and see if we can find any correlation.

Finally, we will look at various techniques to mine emotions/opinions from a text and various approaches that can be used to mine text, such as keyword-based, learning-based, and hybrid approaches. The best approach will then be selected to get attitudes from the text; a model will be trained to see whether researchers' claims regarding older people's attitudes towards technology are in sync with our findings.

The next chapter 2 will deal with the research that has already been undertaken related to the chosen topic and the methods we will be using throughout the research.

2 Literature Review/Background

In this chapter 2 we will have a look at the various research that has already been done in this area and try and make sense of the results to determine any gaps that can be bridged by the research being undertaken by us.

2.1 Older People and Technology

2.1.1 Why is it important?

In the past few decades, we have seen a general trend wherein more people move towards single households and private care systems. On top of that, the cost of traditional health care systems for older people is becoming incomprehensible. This has led to the promotion of technologies and support that can help increase the independence of older people [18].

The pandemic was an eye-opener for everyone, especially for people who did not even meet their loved ones for the last time. It pushed the point of older people being independent further to the extent that it just could not be ignored anymore. Older people were thought to be the most at risk, not just from a physical point of view but also from a mental and emotional point of view also arose due to the social isolation put on everyone because of the containment strategies. The increased risks led to calls being issued worldwide to tackle the dangers associated with the isolation of older people and how they can be diminished using technologies that can help people connect in those testing times [10]. The technologies would help them connect with their close ones and assist them with their daily chores, such as financial planning, keeping track of their health, etc., and help them live longer and more independent lives [30].

Older age is also correlated with a loss of cognitive abilities which can lead to a decline in the capabilities that are instrumental for activities related to their day-to-day lives, such as preparing meals, shopping, doing housework, transferring money, etc. This is where Assistive Technologies (AT) comes into the picture. They compensate for the declining cognitive and decreased motor skills. These can be either low-level, such as ramps or security handles, or high-level, such as safety alarms and monitoring systems [18]. Studies also suggest aging as

a process of selective optimization with compensation, i.e., because of the certain losses that older people experience due to aging, they may optimize resources to increase their capabilities to maintain their goals. For example: instead of using any other mode of transport such as a bus/car, older people may choose to walk or cycle so that they can train their motor abilities, but when even those optimizations are no longer feasible, it can prove to be a challenge to find other solutions. This is where AT can help them by compensating for the reduced capabilities so that older people can maintain their day-to-day activities and their leisure activities.

Yet, one of the biggest obstacles to the development and growth of the aforementioned technological innovations is the aging people's attitude and the willingness to be open to confronting the change they will cause in their everyday lives. In general, it is seen that older people are way behind in adopting new technological advancements, especially compared to the younger generation. However, if the technology adds value to their life, it can be a different ball game altogether [30].

2.1.2 What are the hindrances?

Various literature suggests that older people may be open to using new technologies. However, there are still a few skepticism or hindrances in their mindset or attitudes that they must overcome first. Vaportzis et al. [30], in their study, found the below-mentioned significant themes in the answers when asked to use a tablet: **Barriers to using technology** include a lack of guidance and instruction; for example, the manual may contain technical jargon that is hard to understand for older people. Other barriers include a lack of knowledge and confidence, which is not usually high in older people when asked to use technologies. Cost can also be a significant factor since most smartphones or health devices today are on the pricier side of goods. Health also plays a notable role in whether or not older people can use or accept technology. **Disadvantages of using technology** consisted of the issues because people were not encouraged to use the tablets, felt that it was too complex, and made them feel even more inadequate compared to younger generations. They also thought that the over-usage of tablets could lead to decreased real-life social interactions and communications. These themes were common regardless of whether people had prior experience using computers/technology. However, this study was done before a pandemic hit the world, and a lot has changed since then concerning how everyone, not just older people, views technology.

Another study was done by Bian et al. [1], wherein they wanted to study the perspectives of the older generation towards technology and wearable technology to assess frailty in people in their home settings. They found that most participants generally had a positive outlook toward technology and were open to installing health monitoring devices in their households other than the regular camera due to security and privacy. Even though they raised concerns about appearance, privacy, and the sheer adherence to wearing the devices, the overall attitude

was very positive. However, an interesting observation is that the mean number of years of education for the focus group members was 15, with just one under 10. It is well-known that education is positively correlated to acceptance of new technology. Another important aspect is that the third meeting of the focus group was done using teams due to social containment norms during the pandemic. Thus, most of the members were either already well versed in using technology or were learning to use it.

A similar study was done by Nicole et al. in 2019 [11] on the attitudes of elderly Austrians towards new technologies where used data from the Survey of Health, Ageing and Retirement Europe (SHARE) to judge the attitudes of people towards technological advancement in health and support versus communication and entertainment. They showcased that even though people's preferences towards the system depended on gender, males valued communication and entertainment-based technology, whereas women preferred technology targeted towards health and support. The overall attitude towards new technology was positive, with the slightest interest being towards social media in the older generation, with a net positive score of just 29%. In contrast, alarm systems enjoyed the highest interest from elder adults at around 70%.

In their study, Harris et al. [12] found that the significant barriers that older people had towards accepting technology across three different innovative technologies were mainly due to ignorance of the features of that technology. Most of the products made today do not just fulfil a single purpose; it might be hard for older adults to understand all the features entirely. Moreover, the price point of the technologies in question as innovative technologies may not be at the top of the list of priorities for older people who at times rely on the money they get from retirement or government schemes.

The main issue with all the studies is that they all revolved around a few selected people, formally known as a focus group, which does not give a complete picture. For example, the study by Bain et al. just involved 15 people, and the one by Vaportzis et al. constituted just 18 people, and the results might likely be skewed. Today, we can get views from a much wider audience more efficiently using social media and other online network groups.

2.1.3 What are the facilitators?

Regardless of the hindrances mentioned above, the acceptance of technology amongst the older generation is still on the rise, and few factors are helping. In their study, Hasse et al. [10] found that more than half of the responders have adopted some form of new technology since COVID, and apart from the necessity of the pandemic several, several other factors were influential facilitators for the same. The respondents believed that it would have been much easier for them to adapt to new technologies if they had any prior knowledge or familiarity with technology. For instance, some of the responders were going online and using the web

to learn about technology and stuff they were confused about, and it is easier for people with prior knowledge. Many people also said they were heavily dependent on their friends and family to use any technology since they had better knowledge about it. This support access was also a significant facilitator in adapting to new technology. Older grown-ups also desired that it would be much better if these technologies came up with better instructions in case they were stuck somewhere. Apart from that, societal factors play a significant role as facilitators and hindrances since human beings are social animals and what society thinks about us affect us enormously.

Similarly, Valportiz et al. [30] concluded from their surveys that the significant advantages that older people saw for using tablets were the features of the tablet, the ease with which they could access those features and the personal will they had to adopt new technologies in general. Those features included the large and clear screen and the ability to use it for several things, such as calling friends and family, tracking their health, and taking photos. Giving them the ability to access the information instantly at a click also proved to be an essential facilitator for older adults. Again, social aspects such as the feeling of being left out or not keeping up with their peers also became a vital factor.

Similarly, Bian et al. [1] also suggested that the overall response of all the participants towards the suggested technologies in their study was somewhat optimistic, apart from the support for the standard camera. They will be willing to install some of them in their homes as well since they could get help from the technology in preventing adverse outcomes for specific problems by early detection, and they will be able to live better and longer lives. They thought that the technologies would not only help them with their daily lives but could also suggest them adjust their lifestyles towards a healthier one. The older people went as far as to say that the data collected could also help clinicians to detect several problems in their nascent stages. People also thought that with the help of those technologies, they would not even have to go to care centres, and the technologies would promote care in place.

Furthermore, Haris et al. [12] also found that across the models of STAM and STAAM, the most significant facilitators are dependent on the user's perception, whether they feel the technology is valid or not and whether they feel it is easy to use or not. Social factors and facilitating conditions also significantly influenced the intention to use technology amongst older people.

2.2 Focus Groups Versus Social Network Groups

Focus Groups are a method of getting data where a small group of people, often from 10-15, are brought together to share their views about a designated topic in a moderated way. The group comprises people with a specific range of demographic traits and would help to

gain insights on a topic of choice of the concerned research. Focus Groups have been used for market research for decades to learn about people's perceptions about different topics, including granular details. They can be easily directed to see what works and what does not and get exciting research information [27].

However, focus groups also come with their fair share of problems, such as if the demographics of the people are not appropriately studied and taken into consideration, it is always possible to get a skewed opinion of the overall picture. Secondly, getting people under the same roof and discussing a topic in detail can be challenging, and people often ask for compensatory payments. Hence, arranging a focus group is expensive, reducing the number of participants that are taken up for the same. Last but certainly not the least, the moderator plays a significant role in every focus group, and their personal opinions can often reflect on the data obtained from a focus group, leading to inaccurate results. They can unintentionally or knowingly lead the discussion to reach a particular conclusion or assumption and discourage people from revealing their honest opinions [32].

Nevertheless, with the internet revolution, people's communication has changed dramatically. They no longer require telephones or letters to connect with their loved ones. They can do so simply with a mouse click or on their smartphones using social media websites such as Facebook, Twitter, Reddit, etc. Furthermore, the best part is that the data from these social media websites, if collected and analysed correctly, can give excellent and valuable insights into the users' perspectives. It is impartial, real-time data, and even though people might not directly tweet/write about something, there are online communities, comments, blogs, and so many other platforms where people might talk about it. The data from these websites can provide information that would have been impossible to get using any focus group regardless of the size and the amount of money spent.

However, collecting the data from focus groups is easy and can be done using either pen and paper or using easy tools such as Google Forms for surveys. On the other hand, extracting data from websites such as Twitter or Reddit is not straightforward and can be a handy task at times. The simplest way is to use off-the-shelf tools such as Import.io and ParseHub. However, they usually require paying and do not provide the functionalities one can get using a web scraper such as Selenium or BeautifulSoup. However, to use them, one must know how to code.

2.3 Web Scraping

The amount of data that is now available online is nearly limitless. Web Scraping is a technique used to extract that data from the internet, and it is done by writing scripts that automatically send requests to the designated web servers and then parse the data that is sent back from

them [26]. It is a way of simulating how human beings interact with websites and collecting the data shown on the website. Many websites provide APIs (Application Programming Interfaces) wherein most of the functions the researchers require are already implemented. The researchers have to use those functions to fetch the data. However, there are times when such APIs are not present or are not enough to fetch the desired data. That is when researchers have to look towards tools such as BeautifulSoup, Scarpy, Selenium etc.

Scarping usually helps to convert the unstructured data that is typically present on the web to more formatted and structured data such as spreadsheets, CSV etc., which can be used to analyse the data. Now, to be able to utilize the tools mentioned earlier, it is necessary that one is familiar with some programming language and python is the most commonly used and most accessible coding language that can be learnt today. Python has various libraries that allow users to utilize the abovementioned tools and scrape data off the web. However, it is essential to understand the differences between all the tools and which can be used in what situation.

2.3.1 Scrapy

Scrapy is a collaborative and open-source framework used to extract data from various websites. It is speedy and is one of the most powerful libraries for data extraction available today. The clear advantage of Scrapy is that it is based on Twisted, an asynchronous framework. It does not send all the requests simultaneously, which a server can assume as a malicious user and block all the requests altogether, preventing the blocking of I/O requests by the server. The other advantages of Scarpy are as follows: It also has inbuilt support for data extracted from HTML sources using both XPath (path of the node in an XML document) and CSS (Cascading Style Sheets) selectors. It runs seamlessly on various platforms like Mac, Windows, Linux etc. It is faster than almost all data extraction libraries and requires significantly less CPU usage and memory. On the other hand, even though Scrapy has outstanding developer community support, the documentation is not well written, and this makes Scrapy not easy to use for beginners [19].

2.3.2 BeautifulSoup

BeautifulSoup is a beautiful tool when we talk about web extraction, as the name suggests. This is because of the ease of use which is one of its core features, and the simplicity with which it can extract data from a web server is terrific. However, the issue with BeautifulSoup is that it requires other modules. It requires a module to send requests to the server and a different module to parse the received data into a structured format. However, the ease with which all this can be achieved makes this tool the best for beginners with a shallow learning curve and extensive documentation with equally impressive community support [19].

2.3.3 Selenium

Selenium, on its own, is essentially a testing tool designed to automate Web applications' testing process and minimize human intervention [19]. It simulates human behaviour and interacts with the browser like a human, and since it was made to test, it can do much more than extract data from web pages. For example, it can be used to click buttons like next or last in the search results and even wait for complete page loads, which is impossible with any other library. On top of that, its ease of use for beginners makes Selenium one of the most powerful tools for web extraction.

2.3.4 Which one to use?

The choice of the tool that must be used comes down to various factors, and the situations in which the tool is being used since each has its own sets of problems and strengths.

- Suppose one wants to add their functionalities to the already provided ones. In that case, Scrapy will be the best option, and it is also beneficial for complex and big projects since it can be easily migrated from existing code bases to new ones.
- If one is a beginner and the project is relatively low-level, then excellent results can be achieved using BeautifulSoup.
- If the speed and time required for data scraping are essential, then Scrapy might be the right choice since it is the fastest library out there because of its asynchronous I/O calls.
- If one is new to programming and has no prior experience, then BeautifulSoup is the way to go. It is the easiest to implement, and with some multiprocessing, one can even overcome its speed constraint.
- If one requires to do a complex task with the browser, such as clicking buttons and browsing through various pages, then selenium might be the right choice since it is the best for automation and is just a tad bit slower than Scrapy.

Nonetheless, before the extracted data can be analysed, it is essential that it is cleaned, lemmatized, and then parsed through a POS (Part of Speech) tagger.

2.4 Part-of-Speech Tagging and Techniques

A part of speech tag is a tag that is assigned to every word(token) in a text corpus to showcase to which part of speech that token belongs and also other grammatical categories such as tense, number etc. POS tagging might not seem that important even though it is the base for many NLP applications such as speech recognition, ML, question answering, processing of

information, disambiguation of word sense, etc. [14]. The POS tags not only make automated text processing possible but also enable the use of linguistic criteria. For instance, it helps in distinguishing whether the occurrence of a word is as a verb or as a noun. In a language such as English, a word can be used as both a verb and a noun; the tagger takes the definition of the word and the context into consideration to decide what POS that token belongs to.

There are primarily two ways in which POS tagging is done, similar to any other Machine Learning model, i.e., Supervised and Unsupervised.

Supervised POS tagging models require an annotated corpus to train the tagger regarding various things, such as the word set, word-tag frequencies, and rule sets. The fundamental principle for their working is probability or frequency; in case of ambiguity, the token is assigned the tag given the most during the training dataset. Supervised models require pre-trained models to learn information regarding the word set, and hence increase in the size of the corpora increases the accuracy of these models [16]. These models are based on an N-gram approach, meaning the tag of the token is based on the tags of the previous n-1 token, which acts as its main advantage and disadvantage. These methods retrieve the correct token in sequences; hence, if any of the tokens in the sentence has a wrong tag, the whole sequence will have wrong tags [14].

Unsupervised POS taggers do not require any pre-tagged corpora. They are based on advanced computing methods such as the Baum Welch, a maximizing expectation algorithm. The tagging is based on rules; the model may look at the linguistic features of the previous token or the next token to decide what tag should be given to the current token. For instance, if the preceding token is an article, the next token will likely be a noun. However, it is tough to define the rules manually; hence, these models are first trained to get a set of rules and then the tagging is done based on those rules [14] [16].

Python provides various libraries that support POS tagging and other NLP tools such as NLTK (Natural Language Toolkit) and spaCy. However, both libraries provide similar tools; Spacy generally has a better repertoire of tools, is more focused on the tasks that need to be done and is aimed at app developers. In contrast, NLTK is more of an exploration tool, helps build something from the ground up and is more suitable for students researchers. Then there are various other taggers, such as Stanford POS tagger, Flair and TreeTagger. Stanford POS Tagger is based on java and is a bit hard to install on python; Flair is one of the most accurate taggers out there; TreeTagger works the best for unknown words; depending on the task at hand, any of them can be used.

After tagging, the last thing left is to analyse the data and find the opinions in the data, which is one of the essential aspects of NLP.

2.5 Opinion Mining of data

Social Media, since its inception, has been an integral part of the lives of people. People not only share their views there but also interact with others regarding their views. These people come from various backgrounds and use different terminologies, and it is hard to analyse the actual opinion behind the data [24].

When talking about older people using technology, none of the research has ever been done on such a scale where people's opinions were analysed from the web. The opinions on the web include not only the opinions of older people on technology but also the opinions of people on older people using technology. Hence it becomes essential that such research is undertaken, and actual unfiltered opinions of the people are judged from the social media websites such as Twitter, Board.ie, Journal.ie and many more.

Emotion detection is the process of detecting emotions from the text. With the advancement in computation and NLP techniques, we can now not only get to know the sentiments of a text, which essentially tells us about the polarity of the text to check if the sentiment is positive, negative, or neutral. We can take it a step further and judge the actual emotion behind the text using various machine learning and analysis techniques along with contextual information. As discussed above, the data extracted from social media websites are generally unstructured and come from people with various backgrounds and cultures; hence, it becomes tough yet vital to check people's emotions regarding a particular topic. It can give us deeper insights into the people's actual feelings and help in taking preventive measures if the opinions are not favourable [24].

There are various ways in which emotion recognition can take place, such as keyword detection, which involves looking for specific keywords in the data and assigning each keyword to a particular emotion word-set, such as happy, sad, or angry. The text is then given an overall score based on how many words from each word set are present. Then comes the lexical affinity approach, which takes the keyword detection approach to the next level and assigns a probabilistic affinity to each word belonging to a particular emotion. However, these assigned probabilities are specific to each corpus. Next is the Learning-based approach, which uses a trained classifier to categorize the data into various emotion classes. The classifiers are already mapped using different machine learning approaches such as K-Means, Support Vector Machines (SVM) and others. Last but certainly not least is the Hybrid approach, where the keyword-based and learning-based methods are combined to give better and more accurate results since it uses classifiers and adds linguistic information from dictionaries to get the results [16]. Most of the research, including the ones carried out by Salam et al. [24] and Chowanda et al. [5], uses the learning-based approach to extract emotions from the text.

In our research, we would use a hybrid approach to detect the emotions from the texts we

have extracted from three influential websites where people, specifically Irish people, share their views: Twitter, Board and Journal. Amongst them, Twitter is one of the major websites. In contrast, Journal is a newspaper website with a comment section where people can freely share their opinion, and Board is a website with various discussion boards anyone can raise. People are free to share their opinion. We would use keywords specific to schemes targeted toward older people, such as Active Retirement Ireland (ARI), SeniorLine and HiDigital. Apart from that, we will also use keywords that we think can give us the opinion of people towards older people, such as Carers or Pensioners. It will be interesting to see what opinions and emotions we can finally mine from the dataset and what the results will be.

The next chapter 3 will deal with all the techniques that were discussed in this chapter, we will go over how each one of them has been used and the reason behind using the same.

3 Methodology

This chapter 3 will consist the overall process of implementation and methodology for the dissertation. We will deal with the techniques used for each section and how were they actually used for the purpose of our thesis.

To understand either the views of older people concerning technology or the views of other people about older people using technology, we needed to find out platforms where people are allowed to share their unfiltered views and have the right to say anything they want to. Hence, Twitter came out as the obvious answer. However, we wanted to go a step further. We started looking for news websites where people were allowed to comment and discussion boards where people freely shared their opinions. We found that The Journal (`https://www.thejournal.ie/`) allows people to comment without any premium subscription on their articles, which was not done by any other newspaper and a website called Boards(`https://www.boards.ie/`) which held discussion boards on almost every topic. Therefore, instead of just using Twitter for the data, we decided to go for these three platforms to get a more expansive and diverse point of view.

3.1 Data Extraction/ Web scraping

The main goal of the data extraction part was not just to extract data but to build a pipeline which others could use to extract the data from the sources mentioned earlier, even if they have no coding knowledge. Hence, the coding was done keeping that in mind wherein we could ask users for various inputs such as the keywords, the date from which and to which they needed the tweets (in the case of Twitter) and the number of articles/discussion boards they wanted to get (in case of Journal and Board).

3.1.1 Twitter

The first part of the project was to get data from these different sources into one place and having no experience in web scraping whatsoever; we started looking for APIs that could be used. Luckily, Twitter has its API that allows one to extract data from Twitter using various

search endpoints, and one can get both current (stream) and historical data. However, a lot has changed after the fiasco of Facebook and Cambridge Analytica.

Twitter no longer gives unlimited access to either tweets or profiles; the access is divided into three levels: Essential, Elevated and Academic. 'Essential' is for developers who are beginners and gives just seven days of data with a limit on the number of tweets that can be extracted per month. 'Elevated' is for companies who want to use Twitter data for research; hence, it gives additional search endpoints and historical data but is paid. 'Academic' is for academics who want to research user behaviour and gives full access to all the data and tweets. At first, we signed up for the Essential access level, but the data was insufficient to perform any analysis. Hence, we tried to sign up for the Academic Research access level, for which Twitter generally sends an e-mail asking for details which include the research purpose, what data will be used and so on. However, a primary requirement is that the researcher's name should be on the institution's website, which would not be possible. Therefore, we started looking for other ways to extract data from Twitter. Then, we came across a python library called snscrape [13], a web scraper for social networking websites like Twitter and Facebook. If used on Twitter, it can scrape everything from user profiles to tweets and hashtags, and it does not require a developer account. Twitter allows us to use such scrapers as long as data is not made public and is just used for research purposes such as sentiment analysis or understanding market trends.

There are two ways of using snscrape, one is by using CLI or command line interface, and the second is using wrapper functions for coding languages such as python. We went ahead with wrapper functions over CLI since they are easy to interact with and can be altered easily in case one wants to do something unique.

The attributes that are available through snscrape are explained in table 3.1.

S.No.	Attribute	Description
1.	url	Permanent Link pointing to the location of the tweet
2.	date	Date on which tweet was created
3.	content	Text content related to the tweet
4.	renderedContent	Appears to be text content of the tweet
5.	Id	Id of the tweet
6.	user	User object containing the following data: username, displayname, id, description, descriptionURLs, verified, created, followersCount, friendsCount, statusesCount, favouritesCount, listedCount, MediaCount, location, protected, linkURL, profileImageURL, profileBannerURL
7.	outlinks	
8.	tcooutlinks	
9.	replyCount	Count of replies
10.	retweetCount	Count of retweet
11.	likeCount	Count of likes
12.	quoteCount	Count of quotes
13.	converstationID	ID of the overall tweet conversation including the replies, if single tweet then tweetID is same as conversationID
14.	lang	Assumed language of the tweet, generated automatically
15.	source	Source of where the tweet was posted from, example: Android, iPhone
16.	media	Media Object containing previewURL, fullURL and type
17.	retweetedTweet	If it is a retweet, then ID of original tweet
18.	quotedTweet	If it is a quoted tweet, then ID of original tweet
19.	mentionedUsers	User objects of any users mentioned in the tweet

Table 3.1: Attributes of Tweets extracted using snscrape

The next part was getting tweets from a geo-location such as Dublin in our case. The Twitter metadata has two classes of location, namely, Tweet location, which is the location of the tweet if the users agree of shares, it and Account location, which is based on the 'home' location provided by the user. We first tried to get data based on the tweet's location but found that nearly 1-2% of the tweets that were made had any location mentioned on them; hence, this idea was dropped early on. We then tried using the account location for scraping tweets by location, but it skipped the tweets with geo-location turned on. We then found that since snscrape uses the Twitter search endpoint, we can use the 'near' parameter, which will give me both the geo-tagged and non-geotagged tweets [29], and we went ahead with that approach.

We also wanted to see the posts of journalists to see what they are tweeting. Since journalists, in general, represent the ideology of the institution that employs them. We segregated the top 100 journalists from the Murray Index into the institutions they were a part of and extracted their tweets to get their position on various topics concerning older people. Murray Tweet Index is published by communications consultation company Murray and measures various journalists across different parameters such as engagement level and content quality to rank them in a systematic order [6].

```
Do you want to search from the list of journalists(Y/N): N
Please enter the keyword you want to search: Carers
Please enter the from date (YYYY-MM-DD): 2020-01-01
Please enter the to date (YYYY-MM-DD): 2022-07-25
Please enter the location (Enter 'None' if not needed): Dublin
Tweets extracted : 0
Tweets extracted : 1
Tweets extracted : 2
Tweets extracted : 3
Tweets extracted : 4
Tweets extracted : 5
Tweets extracted : 6
Tweets extracted : 7
Tweets extracted : 8
Tweets extracted : 9
Tweets extracted : 10
Tweets extracted : 11
Tweets extracted : 12
Tweets extracted : 13
Tweets extracted : 14
```

Figure 3.1: Parameterized approach to Extract Tweets based on Dublin as location

Figure 3.1 shows the python script that takes parameters such as the keyword, the location, the start date, and the end date for the search. The script then extracts tweets from Twitter which contain the given keyword in the given range; it also asks whether the user wants to extract tweets from the list of journalists, which gives the tweets from just those journalists in the given date range.

3.1.2 TheJournal.IE

When it came to Journal.ie, we knew that we could not use any pre-existing APIs since there were not any available for it. Then it was time to decide which web scraper we wanted to go ahead with to scrape the data. Since we were starting our journey with web scraping, we went with BeautifulSoup as it was the easiest to learn and implement, but the first thing to extract the data was to find the id of the HTML elements we wanted to extract. A URL can be divided into two main parts: the base URL and the site-specific location. The base URL is the URL at which the site is hosted, for example, `https://www.thejournal.ie/`; then comes the site-specific location, which is the location for individual resources and usually ends

with .html. It is unique for each resource on the website and hence can be used to extract different parts from a website.

Now, the task was to automate the process of article search. This is usually done using the query parameters used by websites to encode values used to perform a search. For example, if one wants to search for jobs on indeed for the software development role in Ireland, then all they have to do is go to `https://ie.indeed.com/` and type software developer in the search bar. Now, one can notice that the address bar changes from `https://ie.indeed.com/` to `https://ie.indeed.com/jobs?q=software%20developer&l=Ireland`, here everything after ?q= is the query parameter which can be changed to get a search result on anything on the website.

At first, the prominent part that we wanted to extract was the article's content, but it is essential to know the structure of the data inside the HTML response page. This can be done by Inspecting the webpage using the Developer tools; these tools allow one to check the document object model or DOM of the website to understand the structure of the website and its source code. All one has to do is click on an element of the website, and the inspect tool will show the code behind that element. We checked the source code of the article body using the same, and therefore as per Figure 3.2, we went with articleContent HTML element to extract the same.

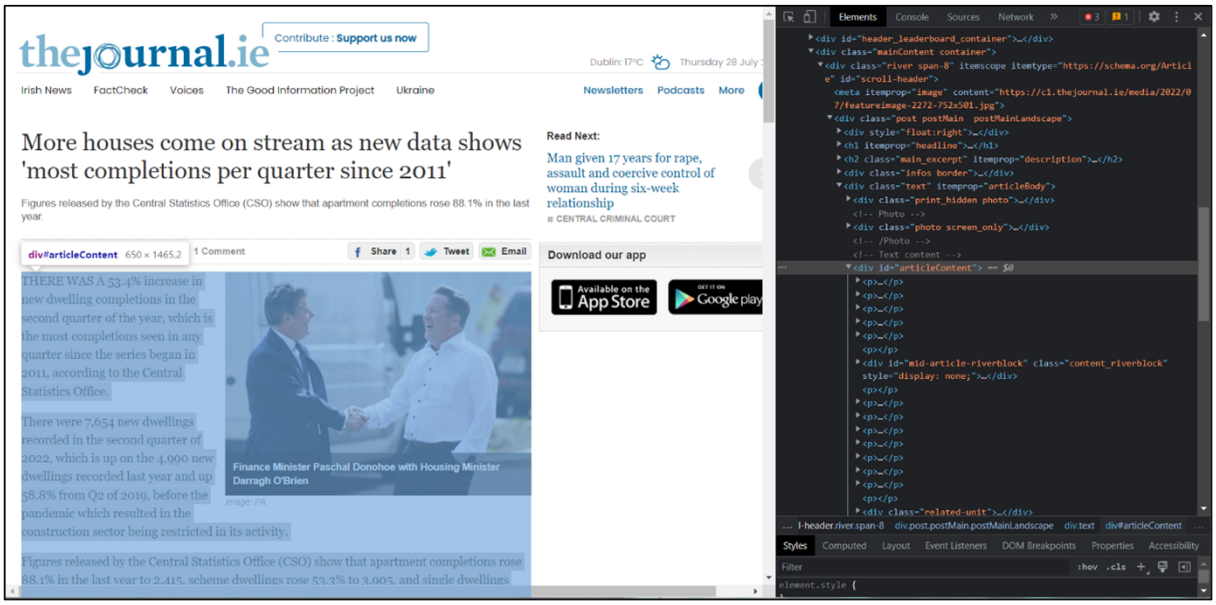

Figure 3.2: HTML Elements from an article of thejounral.ie

Getting the article content was easy, but the problem arose when we tried to get the comments from the article. Websites are generally either static or dynamic. Static websites are ones where the content is already saved on the HTML page, and as the page is loaded, all the data can be accessed. On the other hand, in a dynamic website, none of the content is in HTML

form; instead, a request is sent, and a javascript response completely differs from the HTML source code one sees on the DOM. Therefore, to scrape data from dynamic websites, one has to send the exact requests as done by the browser and then parse the received javascript response.

We found that the comments from the article were being loaded using javascript; thus, we could not use BeautifulSoup to extract those. We then had to switch to Selenium [26] and wait for every article to load completely using the sleep python command for 10 seconds so that the selenium driver could receive the response to the sent request. We then used the XPath of the comment, which is the path of the XML document element, to extract the comment individually from the comment list.

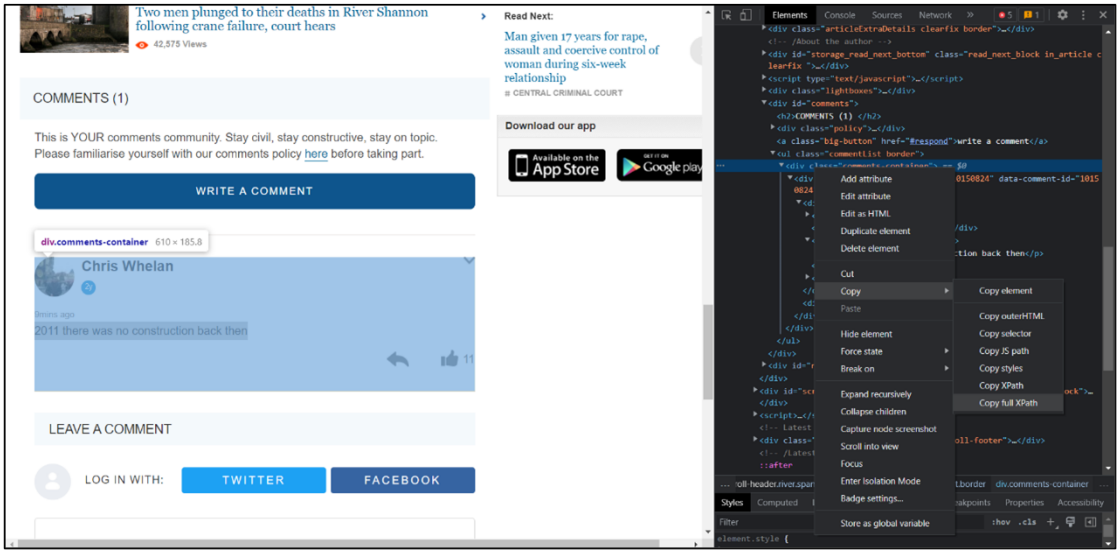

Figure 3.3: Copying XPath from an article

Figure 3.3 shows how XPath can be found and copied from the inspection tab of a web-page.

We then searched using the required keywords and query parameters on thejounral.ie website then copied the 'href' or hypertext reference of every article from the search page and parsed both the content and the comments using an HTML parser.

Figure 3.4 shows the python script for extracting the articles and the comments from thejournal.ie taking 'alone' as the keywords and '5' as the number of articles. The script also tells the total number of comments it extracted.

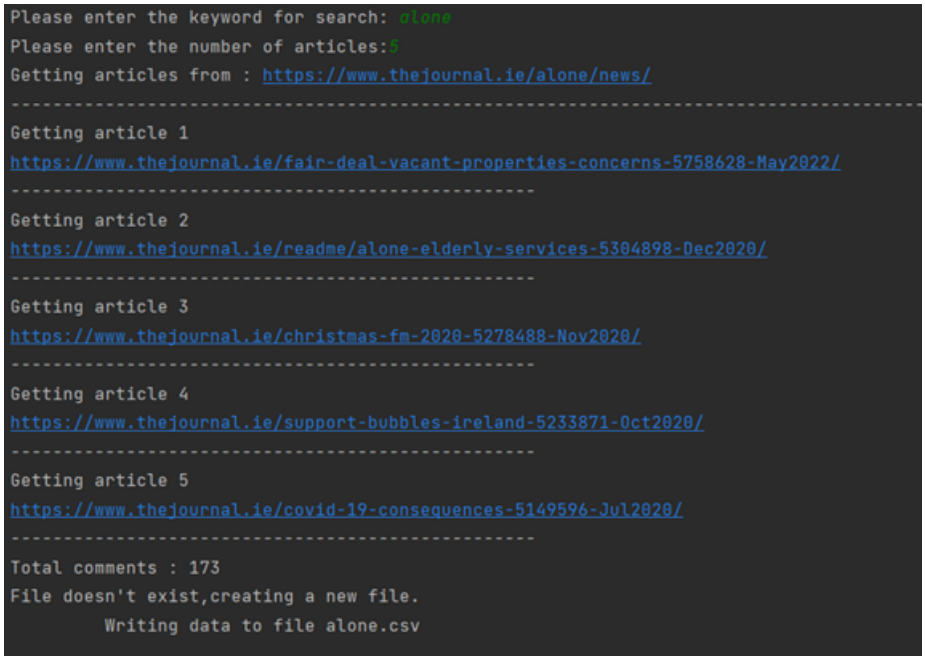

Figure 3.4: Copying XPath from an article

3.1.3 Boards.IE

The process of Baords.ie was pretty similar to the one used for thejournal.ie since it too was using javascript to load the comments, the only difference being that where every search page on thejournal.ie had 40 articles whereas every page on Boards.IE had just ten discussion boards. However, since we were using Selenium, we could quickly unravel the issue by asking the driver to switch to the next page and copy all the href links from all the pages into a single list. Figure 3.5 shows the href links for the discussion boards using 'active retirement' as the keyword.

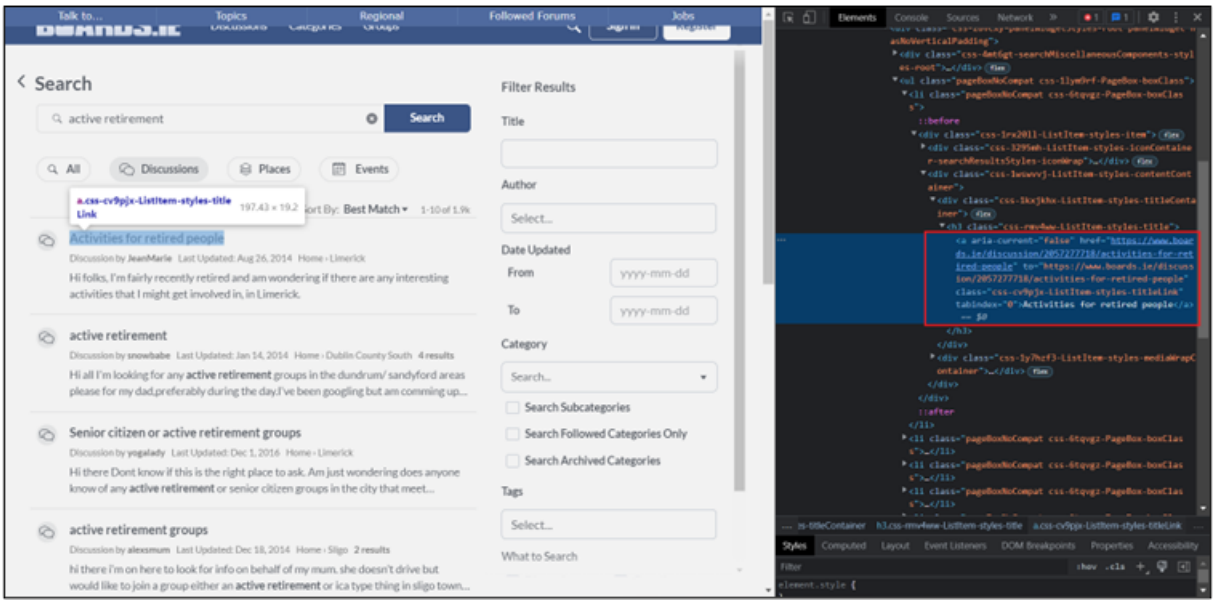

Figure 3.5: href links for various discussion boards on the search page of boards.ie

Figure 3.6 shows the python script for extracting the articles and the comments from boards.ie taking 'active retirement' as the keywords and '5' as the number of discussion boards. It shows the complete list of discussion board's hrefs and extracts all the articles; in the end, since the file with the same name as the keyword already exists, it asks the user to give a new name to the file and saves the data in that file. The script also tells the total number of comments it extracted.

```
Please enter the keyword for search: active retirement
Please enter the number of discussion boards: 5
['https://www.boards.ie/discussion/2057277718/activities-for-retired-people', 'https://www.boards.ie/discussion/2057124621/active-retirement', 'https://www.boards.ie/discussion/205
----------------------------------------------------------------------------------------------------------------------------------------------------------------
Getting article 1 using link https://www.boards.ie/discussion/2057277718/activities-for-retired-people
Discussion Heading: Activities for retired people
----------------------------------------------------------------------------------------------------------------------------------------------------------------
Getting article 2 using link https://www.boards.ie/discussion/2057124621/active-retirement
Discussion Heading: active retirement
----------------------------------------------------------------------------------------------------------------------------------------------------------------
Getting article 3 using link https://www.boards.ie/discussion/2057677213/senior-citizen-or-active-retirement-groups
Discussion Heading: Senior citizen or active retirement groups
----------------------------------------------------------------------------------------------------------------------------------------------------------------
Getting article 4 using link https://www.boards.ie/discussion/2057346980/active-retirement-groups
Discussion Heading: active retirement groups
----------------------------------------------------------------------------------------------------------------------------------------------------------------
Getting article 5 using link https://www.boards.ie/discussion/2057396435/clubs-activities-for-retired-people-in-kildare
Discussion Heading: Clubs/Activities for retired people in Kildare
Total Comments : 34
File already exists.
Please enter a different name for the file: Active Retirement 2
        Writing data to file Active Retirement 2.csv
```

Figure 3.6: Parametrised approach to extract discussion boards and comments from boards.ie

3.2 Data Structuring

After extracting data from the sources, it must be structured into appropriate formats so it can be readied for analysis.

The fields that were extracted from Twitter were as follows:

- Date: The date on which the tweet was made
- Tweet Id: This is the unique ID provided to every tweet by Twitter themselves
- Text: The actual text of the tweet
- Username: Name of the user who made the tweet (Sensitive Information)
- Like Count: Number of likes on the tweet
- Place: Geolocation of the tweet if shared by the user
- Quote Count: Number of times the tweet was quoted
- Retweet Count: Number of times the tweet was retweeted

Each attribute for the tweets was stored in a different field. The fields were then used to construct a data frame later saved as an excel spreadsheet.

Figure 3.7 shows the data extracted fields using the keyword 'Carer', near location 'Dublin' , '2020-01-01' from the date, and '2022-07-25' being the end date.

	A	B	C	D	E	F	G	H	I
		Date	Tweet Id	Text	Username	Like Count	Place	Quote Count	Retweet Count
0		2022-07-24	1.5512E+18	nursing ca	[redacted]	2		0	0
1		2022-07-24	1.5512E+18	@WendyF	[redacted]	1		0	0
2		2022-07-24	1.5511E+18	@Nerilee	[redacted]	1		0	0
3		2022-07-23	1.5509E+18	@butwhys	[redacted]	5	Place(fullName='Dun Laoghaire-F	0	0
4		2022-07-23	1.5508E+18	@Mikeo5	[redacted]	1		0	0
5		2022-07-23	1.5508E+18	Thanks to	[redacted]	8		0	1
6		2022-07-22	1.5504E+18	@eilishba	[redacted]	0		0	0
7		2022-07-22	1.5504E+18	I hate to t	[redacted]	0		0	0
8		2022-07-22	1.5504E+18	Through t	[redacted]	0		0	0
9		2022-07-22	1.5504E+18	Eugene &a	[redacted]	0		0	0
10		2022-07-21	1.5502E+18	This is bril	[redacted]	74		0	28
11		2022-07-21	1.5502E+18	Why do @	[redacted]	3		0	2
12		2022-07-21	1.5502E+18	Looking fo	[redacted]	24		3	8
13		2022-07-21	1.5501E+18	@newstal	[redacted]	0		0	0
14		2022-07-21	1.5501E+18	We have a	[redacted]	1		0	0
15		2022-07-21	1.5501E+18	Become a	[redacted]	1		0	0
16		2022-07-21	1.5501E+18	Carers of	[redacted]	3		0	0

Figure 3.7: Format of the dataset extracted using snscrape from Twitter

For thejounral.ie the data from each article was saved into the below-mentioned attributes:

- Link: Link for each article
- Title: The title of the article
- Content: The content of the article
- Comments: The comments on the article

Figure 3.8 shows the fields of the data extracted using 'alone' as the keyword from thejournal.ie.

S.No	Link	Title	Content	Comments
0	https://www.thejournal.ie	At best unfruitful and at worst a waste': Charities hit ou	REPRESENTING elderly people will raise concerns at Government plans to change the Fair Deal scheme to try and bring vacant properties back onto the rental market, calling it a â€œwasteâ€. There has also been calls for safeguards to be implemented to ensure that older people are not pressured into renting out their homes by family members who pocket the rental income. The	6y May 10th 2022, 12:41 AM My parents house sat idle for years while my father was in a nursing home and I cared for my mother in my house. My mother could not rent it out because my father was in the fair deal and to get 20% of the rental would be a joke and not worth her while. She wouldnâ€™t even break even after the property management fee and gardener would be paid.
1	https://www.thejournal.ie	My2020: 'I am so angry at those who have breached the	As the country faces further	Dave Gibson
2	https://www.thejournal.ie	Christmas FM returns to the airwaves this weekend	CHRISTMAS FM WILL return to	Darren Mc Mahon
3	https://www.thejournal.ie	NPHET to consider guidance for people who live alone c	Updated Oct 15th 2020, 8:35	Gwen Langford
4	https://www.thejournal.ie	'I can't imagine not leaving your home for four months':	THE â€œUNINTENTIONAL	EvieXVI
5	https://www.thejournal.ie	Charity says older people ringing its helpline 'becoming i	IN THE PAST 10 days, 62% of	Charles Coughlan
6	https://www.thejournal.ie	He saw the plight of older people and couldn't stand idl	THIRTY YEARS HAVE passed	Willie Bermingham
7	https://www.thejournal.ie	Elderly and vulnerable reach out as helplines receive 8,0	OVER 5,000 CALLS were made	PV Nevin
8	https://www.thejournal.ie	Govt launches 'befriending' phone-call initiative for olde	OLDER PEOPLE INVOLVED in	Peter
9	https://www.thejournal.ie	Volunteers wanted for Covid-19 test centres and other	VOLUNTEERS ARE NEEDED for	Pat Corrigan

Figure 3.8: Format of the dataset extracted using BeautifulSoup and Selenium from thejournal.ie

The attributes for the discussion boards from Boards have the same fields as those from thejournal. The content tab, instead of the article in case of thejounral now has the description of the discussion board.

3.3 Data Cleaning

The data fetched from all the sources contain much garbage, which can create problems for further analysis, and hence it needs cleaning before it can be passed further for analysis. The text is first tokenized, converted into a smaller substring, and then the garbage data is removed. The garbage values can mention other users, newline characters (/n), links, or any unknown characters. Apart from this, we also need to remove the stop words, which are insignificant words that can hamper the analysis process drastically [24] [5]. Also, comments from the journal and boards contain other insignificant values such as the time and date of the comment or the word "wrote" and characters "»". Since all the comments are in the format "XYZ wrote » ..." along with punctuation marks.

This is done using the python library NLTK. It has various built-in functions that can quickly achieve all these tasks that would otherwise have taken much effort to complete. It also has a list of stop words and punctuations for various languages, and one can even add words to the list of stop words so they can be removed from the overall text. After removing all the stop words, we can also stem the data; stemming is the process of reducing words to their base forms and is a type of text normalization technique. This, too, can be done using the NLTK library using various algorithms such as Porter Stemmer, Snowball Stemmer, and

Lancaster Stemmer. Out of the above-mentioned, Snowball and Porter are more widely used since Lancaster is a more aggressive stemmer and can leak information where the information retained by the stemmed text is less than the original text. After this, the data is clean and ready for further processing.

3.4 Word Level Statistics, PCA and POS tagging

After cleaning the data, the word level statistics are calculated, which gives the total number of words in the comments for each subject, the total number of unique words, and the average count of unique words. Then the five most frequent words from the comments are found to check what words people use the most. All this is done to get an overall idea of the people's language and their opinions. Then we ran a principal component analysis (PCA) of the text. PCA of a dataset is done to find the principal components of the data or those components that show the maximum variance. It is usually done when a dataset has too many features, and we need to find just a few features that can explain the variance in the dataset in the best way possible. In terms of text, PCA can be used to show those words that can explain the maximum variance of the text or showcase those words used the most.

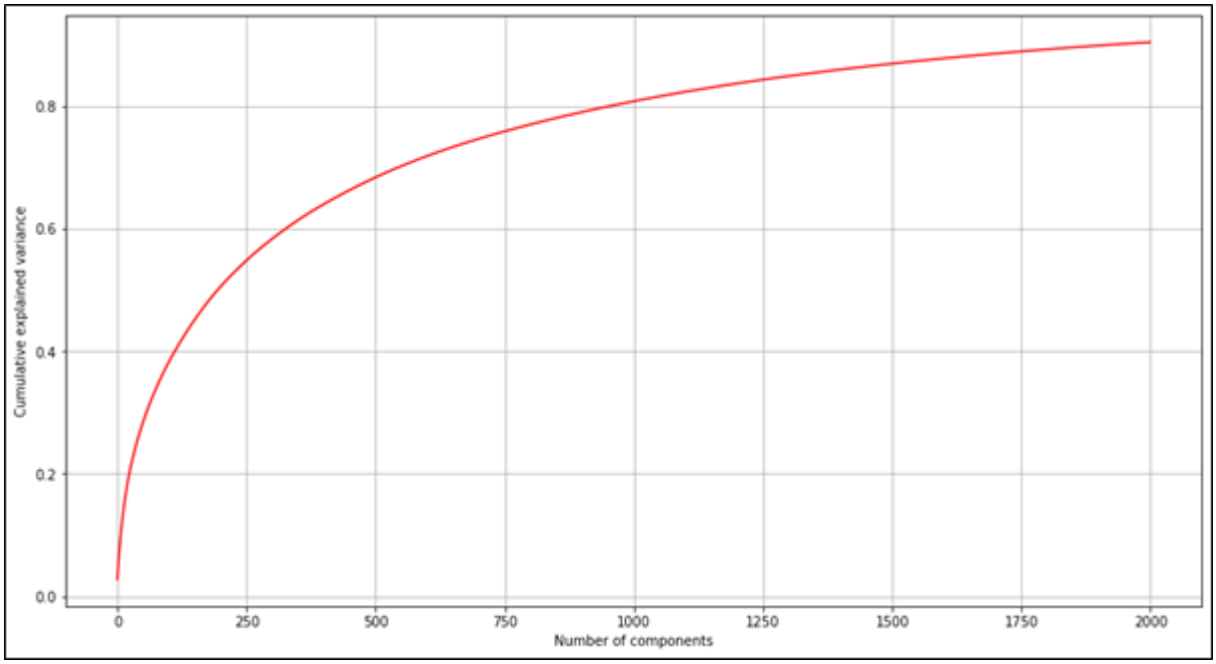

Figure 3.9: Graph showing variance explained vs Number of Components

Figure 3.9 shows a graph showing what variance level is explained by the components. The number of components chosen is based on the level of variance we need to explain from those components; in the above figure, we can choose between 500-750 components since they explain up to 65%-75% of the variance in the dataset. After PCA, we get an idea of the necessary components, and then the text is sent for the POS tagging.

Out of all the taggers available, we have gone for TreeTagger to get the POS tagging done. TreeTagger uses a decision tree to find the transition probabilities for sparse data or data it has not seen. This is the case in our dataset since the language used on Twitter and other online communities are very dynamic. It solves the issue of sparse data by automatically choosing the appropriate size of the contextual information it needs to predict the transition probabilities. It works even better than the state-of-the-art trigram model [17].

Around 55 Part of Speech tags can be assigned by TreeTagger to a token and are part of its tag set. They are explained in Table 3.2.

POS Tag	Description	Example
CC	coordinating conjunction	and
CD	cardinal number	1, one
CDZ	possesive pronoun	one's
DT	determiner	the
EX	existential there	there is
FW	foreign word	d'hoevre
IN	preposition, subordinating conjunction	in, of, like
IN/that	that as subordinator	that
JJ	adjective	green
JJR	adjective, comparative	greener
JJS	adjective, superlative	greenest
LS	list marker	1)
MD	modal	could, will
NN	noun, singular or mass	table
NNS	noun plural	tables
NNSZ	possessive noun plural	people's, women's
NNZ	possessive noun, singular or mass	year's, world's
NP	proper noun, singular	John
NPS	proper noun, plural	Vikings
NPSZ	possesive proper noun, plural	Boys', Workers'
NPZ	possesive noun, singular	Britain's, God's
PDT	predeterminer	both the boys
PP	personal pronoun	I, he, it
PPZ	possessive pronoun	my, his
RB	adverb	however, usually, naturally, here
RBR	adverb, comparative	better
RBS	adverb, superlative	best
RP	particle	give up

POS Tag	Description	Example
SENT	Sentence-break punctuation	. ! ?
SYM	Symbol	/ [= *
TO	infinitive 'to'	togo
UH	interjection	uhhuhhuhh
VB	verb be, base form	be
VBD	verb be, past tense	was, were
VBG	verb be, gerund/present participle	being
VBN	verb be, past participle	been
VBP	verb be, present, non-3rd person	am, are
VBZ	verb be, 3rd person sing. present	is
VH	verb have, base form	have
VHD	verb have, past tense	had
VHG	verb have, gerund/present participle	having
VHN	verb have, past participle	had
VHP	verb have, sing. present, non-3d	have
VHZ	verb have, 3rd person sing. present	has
VV	verb, base form	take
VVD	verb, past tense	took
VVG	verb, gerund/present participle	taking
VVN	verb, past participle	taken
VVP	verb, present, not 3rd person	take
VVZ	verb, 3rd person sing. present	takes
WDT	wh-determiner	which
WP	wh-pronoun	who, what
WPZ	possessive wh-pronoun	whose
WRB	wh-abverb	where, when
Z	possessive ending	's

Table 3.2: POS Tags generated by TreeTagger [25]

Figure 3.10 gives an example of a decision tree used by TreeTagger to determine the tag of the next token based on previous tags [17].

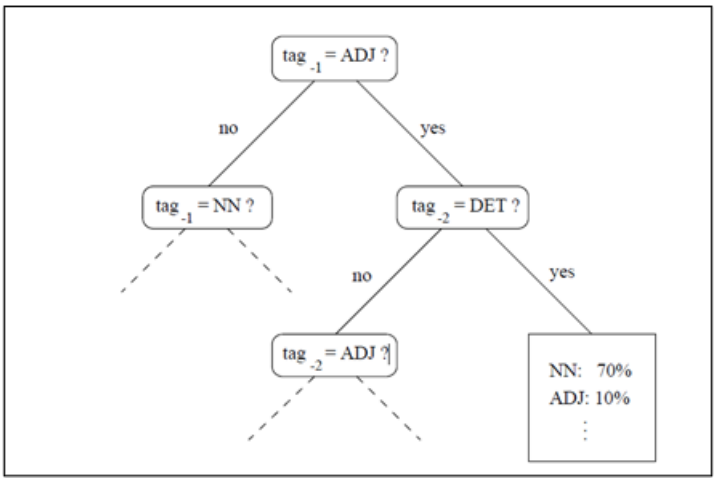

Figure 3.10: Example of decision tree used by TreeTagger

POS tagging helps us to know which part of speech was heavily used by the users in the comments. It is one of the fundamental building blocks for NLP pipelines used for emotion mining and sentimental analysis. In their research, Salam et al. [24] used it to detect unigram, bigrams, and multi-gram; for example, negation followed by an emotion gives the opposite effect (not happy, not scared).

3.5 Emotion Mining

For emotion mining, we took inspiration from the approach suggested by Salam et al. [15], which was a learning-based approach but used just five categories of emotions, namely: Happy, Angry, Sad, Surprise, and Fear, to that I have also added keywords-based approach where each keyword is associated with a particular emotion of the five mentioned above, making my approach a hybrid approach instead of the just learning based one mentioned above. However, before discussing any implemented models, let us first understand how various learning models work.

3.5.1 Logistic Regression

Wikipedia defines Logistic Regression as a mechanism to measure the association between one or more independent variables and a categorical independent variable using a logistic function/sigmoid function. The logistic function is defined using the below-mentioned equation 3.1:

$$\sigma(z) = 1/1 + \exp(-z) \tag{3.1}$$

The hypothesis for Logistic Regression is given by 3.2 :

$$\sigma(z) = \sigma(z)(\beta_0 + \beta_1 x) \tag{3.2}$$

where β_1 is the value for the parameter for the independent variable x and z is the dependent variable.

The logistic regression aims to minimize a cost function so that the function can effectively separate values into different classes. The equation 3.3 gives the cost function for Logistic Regression :

$$J\theta = 1/m \sum [y^i \log(h_\theta(x^i)) + (1 - y^i) log(1 - h_\theta(x^i))] \tag{3.3}$$

This cost function is then minimized using different techniques such as gradient descent, heavy ball, Adam and so on where at each step value of the parameters for the independent variables is changed using the equation 3.4:

$$\theta_j = \theta_{j-1} - \alpha \frac{\partial}{\partial \theta_j} J(\theta) \tag{3.4}$$

where α is the learning rate.

Figure 3.11 shows how the change in parameter values for variables θ_0 and θ_1 effects the value of the overall cost function and how the minimum value of the cost function is reached.

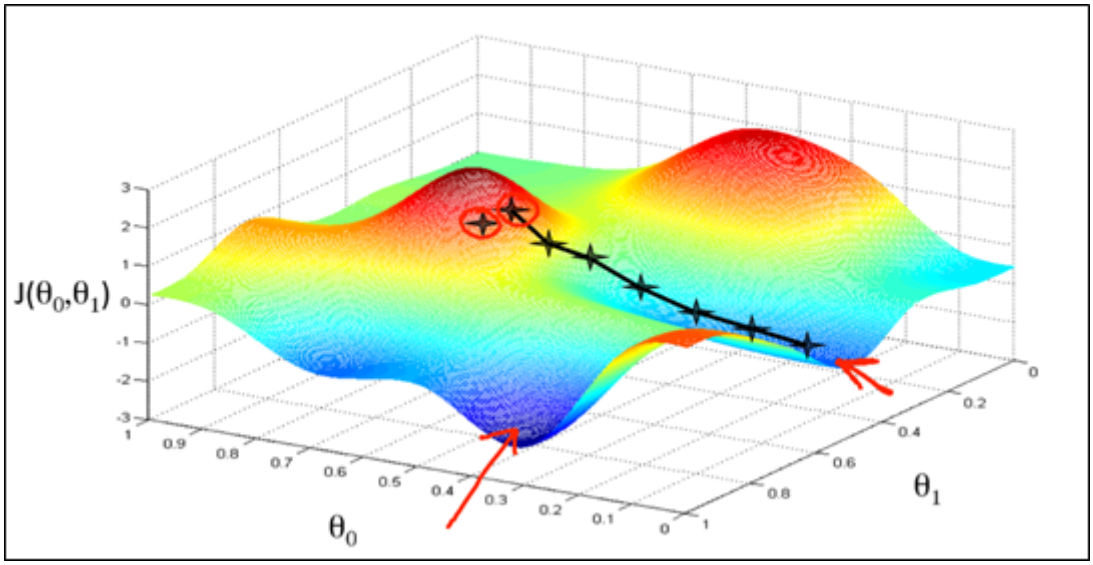

Figure 3.11: Gradient Descent in Logistic Regression [2]

3.5.2 Decision Tree

Decision Trees classify the data into different classes by splitting the complete dataset into smaller subsets so that only the values remaining in the same class are the only ones. The tree is broken down using a process called binary recursive splitting, and at the same time, the overall decision tree is incremented with the new branch/partition.

Binary Recursive split uses a cost function and all parameters at different split points to find the split that gives the least cost; hence it is also known as a greedy algorithm. The Gini index is often used to check how good a split was. It is done by checking how mixed the classes created after the split are, and the purer the classes, the better the split is considered.

The use of binary recursion and the greedy approach of selecting the least cost at all times makes the decision tree much better as a classifier than a logistic regressor.

Figure 3.12 gives an example of a classic decision tree problem along with the splits taken from the scikit-learn documentation.

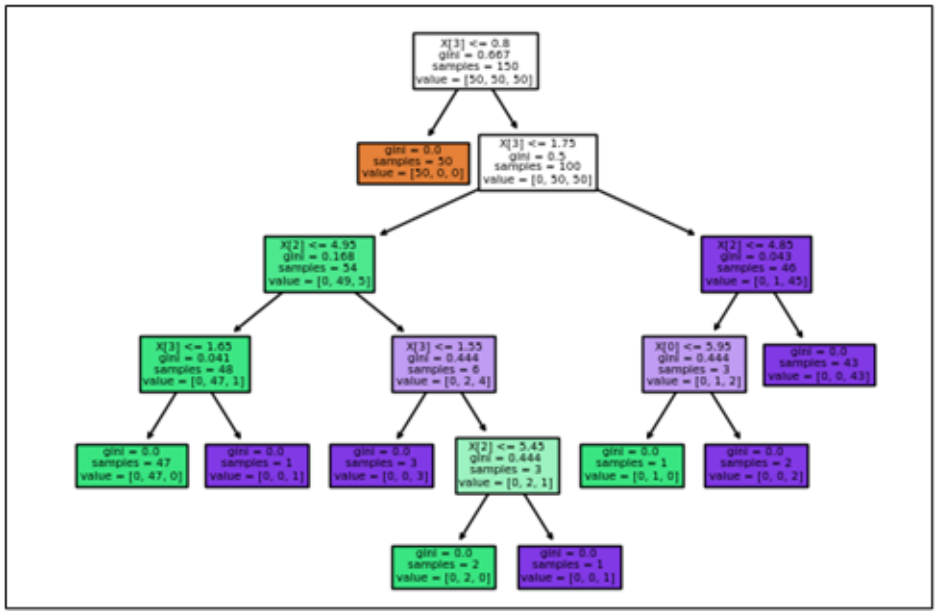

Figure 3.12: Decision Tree on IRIS Dataset

3.5.3 XGBoost

XGBoost for eXtreme Gradient Boost uses gradient boosting algorithms to improve the prediction accuracy of decision trees. Boosting is an ensemble technique where information from several low-level learners is used to improve the overall prediction accuracy of the model. Boosting adds new models that work on the data values predicted wrong by the current model. Whenever a model gets the class for a set of values wrong, its weight is increased so that it is correctly predicted the next time.

Gradient Boosting uses gradient descent techniques to minimize the loss while adding new models to the current models such that the overall loss is lessened, and the model makes the best predictions possible. Ensemble classifiers perform even better than Decision trees, which can be prone to issues such as lower accuracy and higher variance.

Figure 3.13 shows how boosting works for a particular dataset.

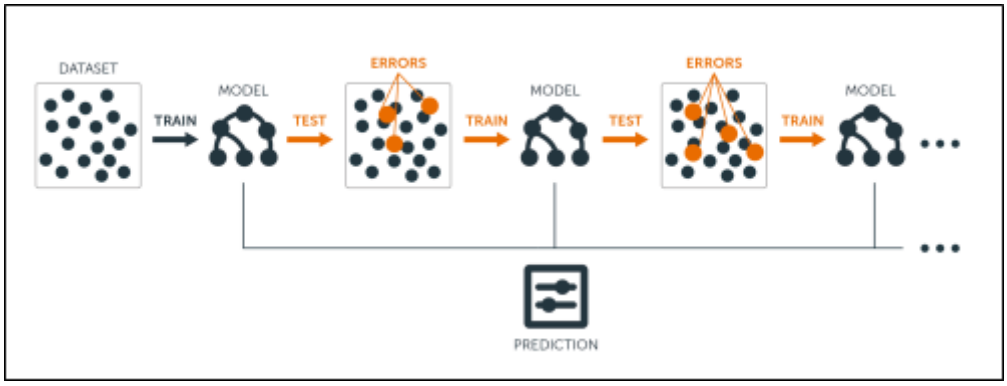

Figure 3.13: The working of Boosting

For our implementation, different classifiers were trained using a Kaggle dataset, which manually categorized tweets into various emotions [28]. We tried various approaches to get good precision; the first was using word counts as features. This was done on the assumption that tweets written when someone is angry or sad are generally longer than ones made when someone is happy. It proved relatively ineffective since all the classifiers (Logistic, DecisionTree) gave an accuracy of just around 30%. The next idea was to use Count Vectorizer, where the complete text is converted into a vector, and each word is replaced by its term frequency [9]; it, too, did not do that well but improved the classifier's performance to 35%. This can be incorporated into the fact that the Count Vectorizer method does not consider the context of the text and just works on the term frequencies. Nevertheless, when the count vectorizer was used with RandomForest and XGBoost, the accuracy increased drastically to over 80%. The last choice was to use the TF-IDF vectorizer, which considers not just the term frequency but also the inverse document frequency and hence the name; this gave the best scores of around 83% using both RandomForest and XGBoost.

Figure 3.14 gives the results from the classifiers using different techniques:

```
Using Word Count and Decision Tree : 0.30667035704400775
Using Word Count and Logistic Regression : 0.3174647107666759
Using Count Vectorizer and XGB Boost  : 0.8339330196512593
Using TF-IDF and XGB Boost: 0.808746194298367
```

Figure 3.14: Accuracy for Different Classifiers on the Kaggle Twitter Dataset

Then, to get a hybrid emotion recognition model, we also associated all the emojis with an emotion so that we could measure the emotions from there, taking inspiration from Vogel et al. [31]. However, we used over 240 emotions instead of a few emoticons. The occurrence of each keyword or emoticon added value to the overall score of that emotion. Also, we provided the classifier with various terms that can be used when people try to portray different emotions, for example, 'elated' when they are happy or 'devasted' when they are sad, to give the classification some linguistic information as well.

```
Dataset\Active retirement Board.csv
[nltk_data]   Package wordnet is already up-to-date!
Counter({'NN': 6271, 'NNP': 2823, 'JJ': 2508, 'VBP': 697, 'IN': 493, 'RB': 422, 'POS': 299,
Total Number of words: 53906
Total Number of unique words aka vocabulary size: 82
Average count: 657
word:years, percentage:0.8
word:retirement, percentage:0.67
word:would, percentage:0.63
word:n't, percentage:0.63
word:time, percentage:0.59
{'Happy': 0.16, 'Angry': 0.08, 'Surprise': 0.16, 'Sad': 0.18, 'Fear': 0.41}
```

Figure 3.15: Results for Active Retirement file extracted from Boards.ie

Figure 3.15 shows the result for a particular file after passing through all the stages of the python script. We can see the total number of words, the total number of unique words after removing all the stop words, and the average count of each word. Then we see the five most used words, after which is the overall emotion of all the comments. All five emotions are given a percentage based on the hybrid approach.

The next chapter 4 will deal with the evaluation of the processes discussed in this section wherein we will try and evaluate how successful or unsuccessful those techniques have been.

4 Evaluation

In chapter 4 we will check how successful the processes used in the chapter 3 have been. We will try to evaluate the overall performance and draw sensible reasoning behind the results seen.

The first and central part of evaluating the overall performance of this research project is to evaluate the performance of the data pipeline that we have created to monitor the online chatter about older people. So, various keywords were used on the three websites, and each returned a different number of hits. We will check how many hits we got, and out of them, we will then try and roughly check the precision of the hits we got by manually checking the results we received.

Figure 4.1 shows the keywords used for twitter and the number of tweets that we were able to fetch from each.

Keyword	Nearby Parameter	Tweets
Active Ageing	None	2346
Active Retirement Association	None	99
Active Retirement Ireland	None	111
AgeAction	None	7838
AgeAction	Dublin	127
AgeandOpportunity	None	778
AgeFriendly	Dublin	5
AgeFriendlyIrl	None	4164
AloneIreland	None	67
Carer	Dublin	296
Friend of Elderly	None	20000
Friend of Elderly	Dublin	11
GenerationTech	None	90
HiDigital	None	536
Old People	Dublin	1324
Older People	Dublin	774
Pensioners	Dublin	154
SeniorLine	None	178
Third Age Ireland	None	27

Figure 4.1: Keywords used for fetching data from Twitter

Here, the nearby parameter tells whether any location filtering took place or not; if its value is None, that means the resultant tweets can be from any part of the world, but if the Nearby location is Dublin, it means that the resultant tweets are only from Dublin and nearby areas. As we can see, many keywords were used; out of them, some were related to schemes that are specific to Ireland, and therefore with them, no nearby parameters were used, such as 'AgeFriendlyIrl', which is a public welfare scheme specific for older Irish people whereas for generic terms such as 'OlderPeople' or 'Carer', the nearby parameter was set to Dublin to limit the results. The tweets were then manually checked to find out the accuracy or precision of the data pipeline; for each keyword, only 50 tweets were randomly checked manually, and for keywords where total tweets were less than 50, all of them were checked. Figure 4.2 shows the results for the same:

Keyword	Tweets	Tweets Checked	Related	Unrealted	Accuracy
Active Ageing	2346	50	41	9	82
Active Retirement Association	99	50	37	13	74
Active Retirement Ireland	111	50	46	4	92
AgeAction	7838	50	42	8	84
AgeAction	127	50	36	14	72
AgeandOpportunity	778	50	48	2	96
AgeFriendly	4	4	4	0	100
AgeFriendlyIrl	4164	50	42	8	84
AloneIreland	67	50	43	7	86
Carer	296	50	44	6	88
Friend of Elderly	20000	50	39	11	78
Friend of Elderly	11	11	10	1	90.90909
GenerationTech	90	50	38	12	76
HiDigital	536	50	34	16	68
Old People	1324	50	37	13	74
Older People	774	50	40	10	80
Pensioners	154	50	47	3	94
SeniorLine	178	50	48	2	96
Third Age Ireland	27	27	26	1	96.2963

Figure 4.2: Accuracy for the Data Pipeline on Twitter

The results of the data pipeline were good, and the pipeline fetched accurate data for most of the keywords other than HiDigital, which had an accuracy of just 68% contained tweets from a promotion on some other website's digital platform. The overall average accuracy for the pipeline on Twitter was 84.01%.

Figure 4.3 shows an example of a tweet that is unrelated to the required keyword and might add up the inaccuracy of the dataset extracted.

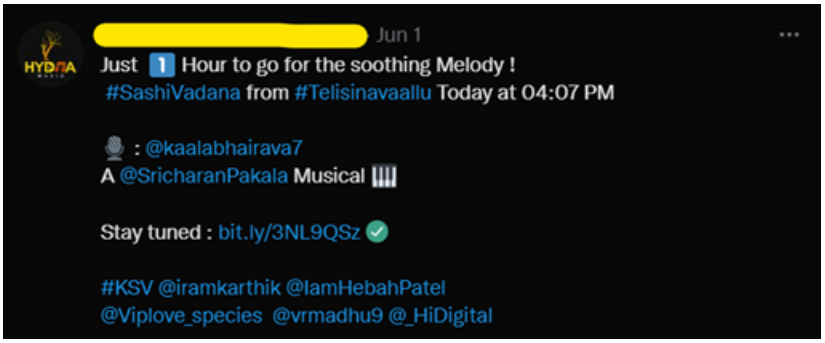

Figure 4.3: Example of an unrelated tweet

Figure 4.4 shows the keywords used to fetch data from TheJounral.ie with the accuracy of the articles fetched from the pipeline.

Keyword	Number of Articles	Total Comments	Articles Checked	Related	Unrealted	Accuracy
Ageism	14	263	10	9	1	90
Alone	40	1089	10	8	2	80
Elderly	40	1097	10	7	3	70
Old People	16	357	10	9	1	90
when-I'm-65	2	312	2	2	0	100

Figure 4.4: Keywords used for TheJounral.IE along with the accuracy

The results for TheJounral.ie were much better and way more accurate than Twitter, showing an average accuracy of around 86% meaning most of the articles fetched were related to the keyword. The pipeline, though, gave the best results for Boards.ie where the accuracy was around 88%, indicating that nearly all the discussions were related to the keyword. Figure 4.5 shows the results and keywords used on Boards.ie.

Keyword	Number of Discussion Boards	Total Comments	Boards Checked	Related	Unrealted	Accuracy
Active retirement	15	196	10	7	3	70
Ageism	10	158	10	8	2	80
Alone elderly	10	71	10	8	2	80
Carer	50	284	10	9	1	90
Elderly	25	202	10	10	0	100
Lifelong Learning	10	36	10	9	1	90
Old people	20	354	10	10	0	100
Older people	20	310	10	9	1	90
Pensioners	35	298	10	10	0	100

Figure 4.5: Keywords used on Boards.ie with accuracy for the results

The dataset was sent through a TreeTagger after cleaning it. TreeTagger is a POS tagger, and basically, it divides the data into separate tokens and gives a part of a speech tag to each token based on the word and the context.

Figures 4.6 and 4.7 show the top ten most often used parts of speech in each dataset formed using all the comments/tweets using the keywords.

Title	Column2.1	Column2.2	Column2.3	Column2.4	Column2.5	Column2.6	Column2.7	Column2.8	Column2.9	Column2.10
Active retirement Board.csv	NN: 6271	NNP: 2823	JJ: 2508	VBP: 697	IN: 493	RB: 422	POS: 299	VBD: 285	CD: 283	VBZ: 254
Ageism board.csv	NN: 8130	JJ: 2905	NNP: 2709	VBP: 962	IN: 612	RB: 605	CD: 418	VBD: 364	VBZ: 342	POS: 292
Ageism_2 journal.csv	NN: 5097	JJ: 2671	NNP: 2495	CD: 1427	VBP: 665	RB: 464	IN: 418	VBD: 362	VBZ: 237	VB: 155
Alone elderly board.csv	NN: 1984	NNP: 784	JJ: 774	VBP: 231	IN: 165	RB: 148	VBZ: 107	VBD: 104	POS: 83	VB: 65
Alone journal.csv	NN: 17357	NNP: 9099	JJ: 8896	CD: 4913	VBP: 2038	RB: 1388	IN: 1366	VBD: 1188	VBZ: 753	VB: 554
Alone_1 journal.csv	NN: 15922	NNP: 8478	JJ: 8145	CD: 4526	VBP: 1857	IN: 1266	RB: 1265	VBD: 1104	VBZ: 693	VB: 501
Carer board.csv	NN: 8159	NNP: 3334	JJ: 3250	VBP: 1201	IN: 590	RB: 556	POS: 360	NNS: 352	VBD: 333	CD: 329
Elderly board.csv	NN: 6960	JJ: 2870	NNP: 2753	VBP: 791	RB: 610	IN: 556	VBD: 357	POS: 309	VB: 262	VBZ: 258
elderly journal.csv	NN: 17202	NNP: 9794	JJ: 8902	CD: 5235	VBP: 2117	RB: 1309	IN: 1293	VBD: 1257	VBZ: 812	VB: 523
Lifelong Learning board.csv	NN: 1461	JJ: 523	NNP: 503	VBP: 198	RB: 113	IN: 98	POS: 83	VBD: 59	VBZ: 58	VB: 52
old people board.csv	NN: 12633	JJ: 4914	NNP: 4908	VBP: 1549	IN: 1097	RB: 855	VBD: 649	VBZ: 577	POS: 514	CD: 500
old people journal.csv	NN: 7906	JJ: 3939	NNP: 3556	CD: 1862	VBP: 969	IN: 656	RB: 606	VBD: 488	VBZ: 310	VB: 229
Older People Board.csv	NN: 10127	JJ: 3835	NNP: 3696	VBP: 1248	IN: 873	RB: 762	CD: 521	VBD: 517	POS: 481	VBZ: 396
Pensioners board.csv	NN: 12156	JJ: 4933	NNP: 4426	VBP: 1383	CD: 1193	IN: 833	RB: 811	VBD: 707	VBZ: 502	POS: 426
when-im-65 journal.csv	NN: 4428	NNP: 2253	JJ: 2239	CD: 1257	VBP: 505	VBD: 328	IN: 320	RB: 218	VBZ: 165	VB: 126
Active Ageing.xlsx	NN: 28726	NNP: 25250	JJ: 14425	VBG: 2663	VBP: 2590	VBD: 1786	IN: 1734	CD: 1424	RB: 1408	VBZ: 1217

Figure 4.6: POS Tagging results – Part 1

Title	Column2.1	Column2.2	Column2.3	Column2.4	Column2.5	Column2.6	Column2.7	Column2.8	Column2.9	Column2.10
Active Retirement Association.xlsx	NNP: 1410	NN: 1209	JJ: 563	VBP: 128	VBD: 106	IN: 96	CD: 94	VBZ: 71	RB: 66	VB: 37
Active Retirement Ireland.xlsx	NNP: 1534	NN: 1146	JJ: 574	VBP: 135	VBD: 92	CD: 77	IN: 73	RB: 58	VBZ: 54	VB: 45
AgeAction.xlsx	NNP: 115552	NN: 84193	JJ: 37119	VBP: 8614	VBD: 5579	IN: 5403	VBZ: 5291	RB: 4934	CD: 3732	VB: 2770
AgeandOpportunity.xlsx	NNP: 9522	NN: 7247	JJ: 3975	CD: 781	VBP: 695	VBD: 630	VB: 532	IN: 419	VBZ: 284	RB: 236
AgeFriendly.xlsx	NNP: 80	NN: 38	JJ: 16	RB: 4	VBD: 3	MD: 2	VB: 2	VBG: 2	VBZ: 2	POS: 1
AgeFriendlyIrl.xlsx	NNP: 69928	NN: 36967	JJ: 17207	VBP: 3611	VBD: 2916	VBZ: 2453	IN: 2437	RB: 2318	CD: 1869	VB: 1229
Age_Action_1.xlsx	NNP: 1941	NN: 1293	JJ: 559	VBP: 143	VBZ: 95	IN: 93	RB: 91	VBD: 80	CD: 45	VBG: 36
AloneIreland.xlsx	NNP: 847	NN: 729	JJ: 358	VBP: 82	CD: 65	VBD: 59	IN: 55	RB: 43	VBZ: 40	VBG: 24
Carer.xlsx	NN: 4902	NNP: 3475	JJ: 1944	VBP: 647	IN: 326	RB: 305	VBD: 280	VBZ: 245	CD: 210	VB: 144
Friend of Elderly.xlsx	NN: 370576	NNP: 230092	JJ: 156775	VBP: 56215	RB: 45521	IN: 27240	VBD: 27070	VBZ: 16708	VB: 13242	CD: 12429
Friend_Of_Elderly_1.xlsx	NN: 155	NNP: 121	JJ: 79	VBP: 26	IN: 15	RB: 14	VBD: 10	VBZ: 8	VB: 8	NNS: 7
GenerationTech.xlsx	NNP: 987	NN: 934	JJ: 418	VBP: 86	IN: 60	VBD: 60	CD: 55	VBZ: 52	RB: 42	VB: 34
HiDigital.xlsx	NNP: 6041	NN: 2830	JJ: 1228	VBP: 258	VBZ: 219	VBD: 195	RB: 163	IN: 161	CD: 124	VB: 85
old people .xlsx	NN: 22384	NNP: 12725	JJ: 9374	VBP: 2596	RB: 1579	IN: 1569	VBD: 1457	VBZ: 1053	CD: 888	VB: 681
Older people.xlsx	NN: 12702	NNP: 7816	JJ: 5610	VBP: 1458	RB: 868	IN: 814	VBD: 743	VBZ: 554	CD: 437	VB: 385
Pensioners.xlsx	NN: 2072	NNP: 1587	JJ: 771	VBP: 207	IN: 163	CD: 140	VBD: 133	RB: 114	VBZ: 89	VB: 53
SeniorLine.xlsx	NNP: 2020	NN: 1925	JJ: 950	CD: 407	VBD: 174	VBP: 172	IN: 138	RB: 129	VBZ: 89	VB: 55
Third Age Ireland_1.xlsx	NNP: 494	NN: 357	JJ: 158	VBP: 39	VBD: 28	RB: 24	VB: 21	CD: 21	IN: 19	VBZ: 17

Figure 4.7: POS Tagging results – Part 2

But not all parts of speech were equally used by people in all of the comments/tweets. Figure 4.8 shows the top 10 parts of speech that were used in their order of use.

Part of Speech	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th
PP (Personal Pronoun)	14	8	0	0	0	0	0	0	0	0
NN (Noun)	12	8	0	0	0	0	0	0	0	0
NNP (Proper noun, singular)	8	10	8	0	0	0	0	0	0	0
JJ (Adjective)	0	8	26	0	0	0	0	0	0	0
VBP (Verb be, present, non-3d persc	0	0	0	24	8	1	0	0	0	0
CD (cardinal number)	0	0	0	8	2	2	4	2	8	3
RB (Adverb)	0	0	0	1	5	12	5	7	3	1
VBG (Verb be, gerund/present partic	0	0	0	1	0	0	0	1	0	2
IN (Preposition, Subordinating Conju	0	0	0	0	10	11	9	2	1	0
VBD (Verb be, Past tense)	0	0	0	0	7	6	8	12	1	0
VBZ (Verb be, 3rd person sing. Prese	0	0	0	0	2	1	2	8	15	5
MD (Modal)	0	0	0	0	0	1	0	0	0	0
POS (Possessive ending)	0	0	0	0	0	0	3	1	3	3
VB (verb be, base form)	0	0	0	0	0	0	3	0	3	19
NNS (Noun plural)	0	0	0	0	0	0	0	1	0	1
Total	**34**	**34**	**34**	**34**	**34**	**34**	**34**	**34**	**34**	**34**

Figure 4.8: Top 10 used parts of speech

The results shown in Figure 22 align with those found by Pennebaker et al. [20] that aging individuals start to use more possessive pronouns, which in our case is indicated by the Personal

Pronouns being the most used part of speech. However, they also contradict the study as we can see that in our case, the verbs used by older people or people talking about older people are more in the present form. In contrast, the study by Pennebaker suggests that with the increase in age, older people tend to use more verbs in the past form. This can be attributed to the fact that people are talking about their problems rather than their experiences in the past which was the case in Pennebaker's study.

To make sure that the distribution for parts of speech was because of the keywords used and not the overall nature of the platform, we found out parts of speech in some unrelated keywords such as politics, weather, and technology on both TheJournal.ie and Boards.ie and the results are showcased in figure 4.9.

Part of Speech	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th
NN (Noun)	4	0	0	0	0	0	0	0	0	0
NNP (Proper noun, singular)	0	4	0	0	0	0	0	0	0	0
JJ (Adjective)	0	0	4	0	0	0	0	0	0	0
VBP (Verb be, present, non-3d person)	0	0	0	1	1	1	0	0	0	0
CD (cardinal number)	0	0	0	2	0	0	0	0	0	0
PP (Personal Pronoun)	0	0	0	1	2	1	0	0	0	0
RB (Adverb)	0	0	0	0	0	2	1	1	0	0
IN (Preposition, Subordinating Conjunct	0	0	0	0	1	0	1	0	0	0
VBZ (Verb be, 3rd person sing. Present)	0	0	0	0	0	0	2	0	2	0
VBD (Verb be, Past tense))	0	0	0	0	0	0	0	2	1	0
POS (Possessive ending)	0	0	0	0	0	0	0	1	0	1
VB (verb be, base form)	0	0	0	0	0	0	0	0	1	3
Total	**4**	**4**	**4**	**4**	**4**	**4**	**4**	**4**	**4**	**4**

Figure 4.9: Top 10 used POS in keywords not related to ageing

As we can see from the above figure when the keywords are not related to aging or elderly, the use of personal pronouns goes down significantly as it is not the top-used part of speech in any of the text datasets, and more nouns are used instead.

Then we checked to see the words used the most by the elderly for each keyword to see if we could see any trend or relation between them and the research done earlier. Figures 4.10 and 4.11 give the ten most used words by the elderly and people talking about the elderly.

Keyword	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th
Active Ageing	ageing	active	activeageing	older	running	healthy	retirement	health	moveitorloseit	us
Active Retirement Association	active	retirement	association	community	member	retired	national	employees	years	great
Active Retirement Ireland	active	retirement	skeptical	older	people	digital	age	members	action	support
AgeAction	ageaction	people	older	alone_ireland	nursing	ageathome	age	nursinghomesire	donnellystepher	need
AgeandOpportunity	ageandopportunity	facebook	movement	minutes	11am	join	us	part	searching	covid19
AgeFriendly	crokepark	age	agefriendly	launch	first	friendly	stadium	lordmayordublin	older	people
AgeFriendlyIrl	agefriendlyirl	support	friendly	people	older	ageaction	healthyireland	alone_ireland	great	hselive
Age_Action_1	ageaction	people	alone_ireland	older	activeirl	dlrcc	agefriendlyirl	elderly	agefriendlydcu	jeanm963
AloneIreland	aloneireland	support	alone	youarenotalone	alone_ireland	people	sad	older	donation	thank
Carer	ipad	help	care	family	home	get	people	time	person	dad
Friend of Elderly	elderly	friends	smarttechnology	people	family	one	covid	get	care	like
Friend_Of_Elderly_1	elderly	friends	friendofelderly	time	family	christmas	people	relatives	maths	born
GenerationTech	generationtech	tech	technology	jpmorgan	volunteer	zoom	scared	girls	female	challenge
HiDigital	chloe_hidigital	_hidigital	flutter	viplove_species	iamhebahpate	jk_dr	ksv	telisinavaallu	iramkarthik	itsactornaresh
old people	people	old	year	like	get	young	fear	think	one	years
Older people	people	socialize	stressed	get	need	many	ireland	age	care	like

Figure 4.10: Top Ten used words– Part 1

Keyword	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th
Pensioners	pensioners	pension	japan	pay	0.98	year	time	tax	people	like
SeniorLine	seniorline	older	people	1800	service	freephone	80	45	10am	thirdageireland
Third Age Ireland_1	age	ireland	third	brady	ceo	aine	people	older	kildare	golfgate
Active retirement Board	years	retirement	would	time	discrimination	disheartned	vs	people	work	like
Ageism board	get	one	work	people	age	experience	years	year	would	m
Ageism_2 journal	2014	people	sorry	2021	injustice	age	get	would	19th	2012
Alone elderly board	would	elderly	get	alone	tv	could	know	need	lock	assistance
Alone journal	2020	people	pm	2018	2015	scared	would	elderly	2017	get
Alone_1 journal	2020	pm	people	2018	fear	would	2015	2017	elderly	one
Carer board	carers	allowance	carer	get	would	benefit	time	applied	work	waiting
Elderly board	elderly	would	people	person	one	drive	get	driving	age	road
elderly journal	care	2018	money	2017	people	recovery	2019	home	would	1st
Lifelong Learning board	m	ve	one	would	food	eating	work	friends	people	life
old people board	people	old	would	like	get	years	need	one	think	older
old people journal	people	2014	3rd	2012	get	2017	would	like	2018	elderly
Older People Board	people	older	would	like	get	one	old	age	re	think
Pensioners board	pension	tax	years	get	pay	would	pensions	state	age	retirement
when-im-65 journal	2020	23rd	pm	pension	22nd	age	people	public	pensions	tax

Figure 4.11: Top Ten used words– Part 1

We see that familiar words such as aging, older, retirement, and so on often appear, suggesting that the data extracted is correct and that it is either older people talking about themselves or others talking about older people. We also see words such as sorry, fear, scared, sad, disheartened, and stressed, which might suggest the mentality of the elderly as opposed to the previous studies, which might have been giving out a narrow view of the overall situation [10] [30] [1]. We also see words such as zoom and smarttechnologies, suggesting that older people are indeed open to using newer technologies as opposed to the stereotypes [1] [11]. There are hashtags such as yourarenotalone, which might be used to motivate people that would have been feeling lonely during the testing times of the COVID outbreak. We also see Carers talking about allowance and the benefits that they should be getting. We also see names of people such as 'Stephen Donnelly' and mentions of 'Lord Mayor Dublin', suggesting how older people still take an interest in politics and policies that majorly influence their lifestyles.

Lastly, we tried to get an overall emotion based on the complete text using the hybrid approach we suggested in the methodology section. Figures 4.12 and 4.13 give the overall percentage of emotions calculated from all the comments/tweets using each keyword.

Keyword	Happy	Angry	Surprise	Sad	Fear	Max
Active Ageing	17%	27%	10%	14%	27%	Angry
Active Retirement Association	16%	5%	10%	11%	5%	Happy
Active Retirement Ireland	22%	11%	18%	13%	11%	Happy
AgeAction	18%	10%	14%	21%	10%	Sad
AgeandOpportunity	41%	27%	10%	10%	27%	Happy
AgeFriendly	21%	36%	21%	14%	36%	Angry
AgeFriendlyIrl	32%	12%	11%	15%	12%	Happy
Age_Action_1	38%	11%	11%	22%	11%	Happy
AloneIreland	5%	16%	21%	35%	16%	Sad
Carer	18%	5%	18%	24%	5%	Sad
Friend of Elderly	31%	5%	16%	21%	5%	Happy
Friend_Of_Elderly_1	25%	0%	17%	31%	0%	Sad
GenerationTech	40%	5%	17%	10%	5%	Happy
HiDigital	39%	5%	15%	16%	5%	Happy
old people	16%	8%	16%	25%	8%	Sad
Older people	16%	9%	12%	24%	9%	Sad

Figure 4.12: Emotions from text – Part 1

Keyword	Happy	Angry	Surprise	Sad	Fear	Max
Pensioners	11%	7%	15%	22%	7%	Sad
SeniorLine	47%	7%	10%	10%	7%	Happy
Third Age Ireland_1	21%	34%	12%	16%	34%	Angry
Active retirement Board	17%	8%	16%	18%	8%	Sad
Ageism board	13%	10%	13%	17%	10%	Sad
Ageism_2 journal	11%	11%	12%	20%	11%	Sad
Alone elderly board	17%	6%	18%	46%	6%	Sad
Alone journal	14%	6%	15%	37%	6%	Sad
Alone_1 journal	14%	6%	16%	26%	6%	Sad
Carer board	9%	49%	20%	19%	49%	Angry
Elderly board	11%	8%	17%	18%	8%	Sad
elderly journal	12%	7%	15%	27%	7%	Sad
Lifelong Learning board	39%	7%	14%	12%	7%	Happy
old people board	11%	9%	15%	23%	9%	Sad
old people journal	46%	6%	13%	22%	6%	Happy
Older People Board	17%	10%	15%	21%	10%	Sad
Pensioners board	7%	52%	13%	6%	52%	Angry
when-im-65 journal	12%	13%	8%	20%	13%	Sad

Figure 4.13: Emotions from text – Part 2

Out of the 34 texts, exactly half (17) had an overall emotion of fear based on the classifier's choice and the words in the text. Ten texts had an overall emotion of happiness. These were majorly for the keywords related to schemes such as HiDigital, AgeAndOppurtunity, GenerationTech, SeniorLine, and others. Four texts were associated with a sad emotion; those were extracted using keywords such as AloneIreland, Alone (TheJournal), Alone (Boards), and four conveyed an angry emotion; these might be related to people who are angry with the situation such as pensioners or carers. None of the texts showcased a surprising emotion.

We also thought it would be attractive to the results on a sentence basis when we take the context out of the comments and judge them individually. Therefore, we went ahead and did

the emotion detection on a sentence basis which is shown in figures 4.14 and 4.15.

Dataset	Total Sentences	Angry	Fear	Happy	Sad	Surprise	EBOS	EBOT	Similarity
Active retirement Board	837	54	406	197	84	96	Fear	Fear	YES
Ageism board	963	80	402	275	115	91	Fear	Fear	YES
Ageism_2 journal	625	59	272	172	76	46	Fear	Fear	YES
Alone elderly board	266	9	139	62	21	35	Fear	Fear	YES
Alone journal	2174	126	471	918	398	261	Happy	Happy	YES
Alone_1 journal	2012	123	441	348	852	248	Sad	Angry	NO
Carer board	1032	38	262	126	463	143	Sad	Happy	NO
Elderly board	1017	49	247	528	93	100	Happy	Happy	YES
elderly journal	2058	422	117	458	831	230	Sad	Sad	YES
Lifelong Learning board	158	4	33	82	17	22	Happy	Fear	NO
old people board	1702	131	320	851	242	158	Happy	Happy	YES
old people journal	911	43	264	138	379	87	Sad	Sad	YES
Older People Board	1428	109	236	796	170	117	Happy	Happy	YES

Figure 4.14: Emotion based on individual sentences – Part 1

Dataset	Total Sentences	Angry	Fear	Happy	Sad	Surprise	EBOS	EBOT	Similarity
Pensioners board	1650	77	459	777	186	151	Happy	Happy	YES
when-im-65 journal	482	57	176	147	59	43	Fear	Fear	YES
Active Ageing	2346	525	835	648	198	140	Fear	Fear	YES
Active Retirement Association	99	2	62	19	7	9	Fear	Fear	YES
Active Retirement Ireland	111	11	32	43	9	16	Happy	Happy	YES
AgeAction	7838	3272	684	1887	1172	823	Angry	Angry	YES
AgeandOpportunity	778	145	82	488	34	29	Happy	Fear	NO
AgeFriendly	4	2	0	2	0	0	Angry	Fear	NO
AgeFriendlyIrl	4164	194	721	2622	349	278	Happy	Fear	NO
Age_Action_1	127	9	29	55	22	12	Happy	Sad	NO
AloneIreland	67	2	5	42	7	11	Happy	Sad	NO
Carer	296	17	91	82	57	49	Fear	Fear	YES
Friend of Elderly	20001	391	4202	10021	3373	2014	Happy	Angry	NO
Friend_Of_Elderly_1	11	0	0	6	5	0	Happy	Fear	NO
GenerationTech	90	2	22	50	8	8	Happy	Fear	NO
HiDigital	536	20	70	361	37	48	Happy	Happy	YES
old people	1324	93	380	339	308	204	Fear	Fear	YES
Older people	774	82	205	239	174	74	Happy	Happy	YES
Pensioners	145	7	50	36	31	21	Fear	Fear	YES
SeniorLine	178	14	43	89	23	9	Happy	Angry	NO
Third Age Ireland_1	28	1	18	6	2	1	Fear	Fear	YES

Figure 4.15: Emotion based on individual sentences – Part 2

In the above figures, EBOS signifies the most prominent emotion based on individual sentences, and EBOT signifies emotions based on the overall text. Similarity tells whether the emotion found using the overall text and, on a sentence, the basis is the same or not. However, this would go against the purpose of doing POS altogether, which suggests that context plays a significant role in signifying the emotions of the text. Therefore, we ran a Chi-square test to check whether the values found using both methods were independent or not; the results for that can be found in table 4.1.

Emotion	Chi – Square Test p value
Happy	0.5829
Sad	0.2779
Angry	0.311
Fear	0.2689
Surprise	0.3361

Table 4.1: Results of the Chi-Square Test for each emotion

As we can see, it is evident from table 4.1 that the results found using both methods are independent of each other since none of the p values is less than the significant value of 0.05. Hence, the null hypothesis cannot be rejected.

However, even though the results found using both techniques are independent, we can see that the results are in sync with the most frequently used words. Most of them depicted either fear or sadness, and some were related to happy and excited emotions. The findings are also partially in contrast to the findings made by most literature [10] [30] [1], where it was stated that the overall outlook of the elderly towards technology, in general, is optimistic. We found that the outlook of older people or people talking about using technology might be positive when it comes to schemes where they feel that older people are being included and made for their benefit. However, in general, the overall emotion is still fear, wherein they are not comfortable with the technology or the situations where they are made to use technology.

Nevertheless, the fact that so many people are coming online to talk about older people or the older people themselves coming over to these websites and giving out their points of view makes us believe that that situation is improving.

The next chapter 5 will give an overall conclusion to the dissertation with a brief discussion of the overall results and highlight the main points of the same.

5 Conclusion

In conclusion, the infrastructure to validate the online chatter successfully gave an average accuracy of over 80% on all three platforms. However, the only keyword for which the accuracy was lower than 70% was 'HiDigital' due to the keyword used by some other websites for promotion. This, too, can be handled in the future using the nearby parameter to filter the results according to the location.

Apart from that, if we look at the POS tagging, we find that possessive pronouns are the most used POS by the elderly and are in conjunction with the research by Pennebaker et al. [20], suggesting that with increasing age, people tend to use more possessive pronouns. This was also proved by finding the distribution of POS tags for keywords unrelated to the elderly. However, some of our findings were against other suggestions in the same research wherein they said that with aging, people tend to move towards using more past tense. However, it is because most of the tweets or comments are people talking about the issues they are facing or the problems they are facing and, therefore, more reliance on the present tense.

There were a variety of emotions that were found in the texts; 50% of the text (17 of 34) were found to have an emotion of fear, followed by an emotion of happiness which was found in 10 texts out of 34. 4 texts showed an overall emotion of sadness and anger. Surprisingly none of the texts showed an emotion of surprise. The results were also confirmed on the sentence basis, even though the results found using both were significantly independent. This was also visible in the most frequently used words, such as sad, scared, and anxious, which showed fear and anxiousness amongst the elderly. The positive emotions were more correlated to inclusive schemes such as AgeAndOppurtunity, SeniorLine, and HiDigital, which are government schemes to educate older people to use technology and provide them with opportunities.

In the end, we must remember that there is still work left to be done when we discuss creating inclusive technologies for the elderly. We need to keep in mind that the current generation of older people come from an age when 'apple' was just some fruit, 'tablet' was something on which the commandments were delivered by Moses or something one used to take when they fell ill or a pad that one used to write on. People spent hours looking at Encyclopaedias and not using Google to get some information. While today we use Alexa to change the channels

on TV, back in the day, it used to be our fathers and their voice command asking us to change the channel instead [22].

Bibliography

[1] Chao Bian, Bing Ye, Anna Hoonakker, and Alex Mihailidis. Attitudes and perspectives of older adults on technologies for assessing frailty in home settings: A focus group study, 12 2020.

[2] Rina Buoy. Gradient descent training with logistic regression, 2019.

[3] CentralStatisticsOffice. CSO statistical publication, Dec 2021.

[4] Ke Chen and Alan Chan. Gerontechnology acceptance by elderly hong kong chinese: A senior technology acceptance model (stam). *Ergonomics*, 57, 03 2014.

[5] Andry Chowanda, Rhio Sutoyo, Meiliana, and Sansiri Tanachutiwat. Exploring text-based emotions recognition machine learning techniques on social media conversation. *Procedia Computer Science*, 179:821–828, 2021. 5th International Conference on Computer Science and Computational Intelligence 2020.

[6] Murray Consultants. Murray tweet index, 2021.

[7] Fred D. Davis, Richard P. Bagozzi, and Paul R. Warshaw. User acceptance of computer technology: A comparison of two theoretical models. *Management Science*, 35(8):982–1003, 1989.

[8] Grace Fox and Regina Connolly. Mobile health technology adoption across generations: Narrowing the digital divide. *Information Systems Journal*, 28, 01 2018.

[9] Alifia Ghantiwala. Emotions using nlp based on classifying text, 2022.

[10] Kristen Haase, Theodore Cosco, Lucy Kervin, Indira Riadi, and Megan O'Connell. Older adults' experiences of technology use for socialization during the covid-19 pandemic: A regionally representative cross-sectional survey (preprint). *JMIR Aging*, 4, 02 2021.

[11] Nicole Halmdienst, Michael Radhuber, and Rudolf Winter-Ebmer. Attitudes of elderly austrians towards new technologies: communication and entertainment versus health and support use. *European Journal of Ageing*, 16, 12 2019.

[12] Maurita Harris, Kenneth Blocker, and Wendy Rogers. Older adults and smart technology: Facilitators and barriers to use. *Frontiers in Computer Science*, 4:835927, 05 2022.

[13] JustAnotherArchivist. snscrape. `https://github.com/JustAnotherArchivist/snscrape`, 2022.

[14] Suvarna G Kanakaraddi and Suvarna S Nandyal. Survey on parts of speech tagger techniques. In *2018 International Conference on Current Trends towards Converging Technologies (ICCTCT)*, pages 1–6, 2018.

[15] Simon Kemp. Digital 2022: Global overview report, 2022.

[16] Deepika Kumawat and Vinesh Jain. Pos tagging approaches: A comparison. *International Journal of Computer Applications*, 118:32–38, 05 2015.

[17] Lluís Màrquez and Horacio Rodríguez. *Part-of-speech tagging using decision trees*, volume 1398, pages 25–36. 04 2006.

[18] Claudia Oppenauer, Barbara Preschl, Karin Kalteis, and Ilse Kryspin-Exner. Technology in old age from a psychological point of view. In Andreas Holzinger, editor, *HCI and Usability for Medicine and Health Care*, pages 133–142, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg.

[19] Sri Manikanta Palakollu. Scrapy vs selenium vs beautiful soup for web scraping, 2019.

[20] James Pennebaker and Lori Stone. Words of wisdom: Language use over the life span. *Journal of personality and social psychology*, 85:291–301, 09 2003.

[21] Gil Press. Cleaning big data: Most time-consuming, least enjoyable data science task, survey says, 2016.

[22] Rachel Pruchno. Technology and Aging: An Evolving Partnership. *The Gerontologist*, 59(1):1–5, 01 2019.

[23] Karen Renaud, Karen, Judy Biljon, and Judy. Predicting technology acceptance and adoption by the elderly: A qualitative study. volume 338, 01 2008.

[24] Shaikh Salam and Rajkumar Gupta. Emotion detection and recognition from text using machine learning. *International Journal of Computer Sciences and Engineering*, 6:341–345, 06 2018.

[25] Beatrice Santorini. Part-of-speech tagging guidelines for the penn treebank project, 1991.

[26] Nuha Abdul Rasheed Sarah Fatima, Shaik Luqmaan. Web scraping with python and selenium. *IOSR Journal of Computer Engineering (IOSR-JCE)*, 23(3):01–05, 2021.

[27] Dominick Soar. Where social media beats focus groups, 2011.

[28] Anjenkya Tirpathi. Emotion classification nlp. Technical Report Version 1, 2021. url https://www.kaggle.com/datasets/anjaneyatripathi/emotion-classification-nlp .

[29] Scott Tomlins. Scraping tweets by location in python using snscrape', 2020.

[30] Ria Vaportzis, Maria Clausen, and Alan Gow. Older adults perceptions of technology and barriers to interacting with tablet computers: A focus group study. *Frontiers in Psychology*, 8:1687, 10 2017.

[31] Carl Vogel and Jerom F. Janssen. Emoticonsciousness. In Anna Esposito, Amir Hussain, Maria Marinaro, and Raffaele Martone, editors, *Multimodal Signals: Cognitive and Algorithmic Issues*, pages 271–287, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg.

[32] Alex Writing. Advantages disadvantages of a focus group, 2019.