Original Paper

Mining Social Media Data to Study the Consequences of Dementia Diagnosis on Caregivers and Relatives

Authors names:

Anika Oellrich¹, George Gkotsis², Richard Dobson¹,³, Tim JP Hubbard⁴ Rina Dutta*²

¹ MRC Social, Genetic & Developmental Psychiatry Centre (SGDP), King's College London, London, SE5 8AF, UK
² Academic Department of Psychological Medicine, Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, SE5 8AF, UK
³ Farr Institute of Health Informatics Research, UCL Institute of Health Informatics, University College London, London, WC1E 6BT, UK
⁴ Department of Medical & Molecular Genetics, King's College London, London, SE1 9RT, UK

*Corresponding Author:

Rina Dutta, King's College London, United Kingdom
rina.dutta@kcl.ac.uk

Senior Clinical Lecturer / Consultant Psychiatrist
Academic Department of Psychological Medicine
3rd floor Weston Education Centre
King's College London | Cutcombe Road | London SE5 9RJ
Tel +44 (0)7904 207378 | Fax +44 (0)207 848 5408 |

Keywords:
social media; Reddit; impact dementia diagnosis; automated analysis

Abstract:

Background: Dementia is a growing public health concern with approximately 50 million people affected worldwide in 2017 and this number is expected to reach more than 131 million by 2050. The toll on caregivers and relatives cannot be underestimated as dementia changes family relationships, leaves people socially isolated, and affects the finances of all those involved.

Objective: The aim of this study was to explore using automated analysis (i) the age and gender of people who post to the social media forum Reddit about dementia diagnoses, (ii) the affected person and their diagnosis, (iii) relevant subreddits authors are posting to, (iv) the types of messages posted and (v) the content of these posts.

Methods: We analysed Reddit posts concerning dementia diagnoses. We used a previously developed text analysis pipeline to determine attributes of the posts as well as their
authors to characterise online communications about dementia diagnoses. The posts were also examined by manual curation for the diagnosis provided and the person affected. Furthermore, we investigated the communities these people engage in and assessed the contents of the posts with an automated topic gathering technique.

Results: Five hundred and thirty five Reddit posts were identified as relevant and further processed. Our results indicate that the majority of posters in our data set are women, and it is mostly close relatives such as parents and grandparents that are mentioned. Both the communities frequented and topics gathered reflect not only the sufferer's diagnosis but also potential outcomes, e.g. hardships experienced by the caregiver or the requirement for legal support. The trends observed from this dataset are consistent with findings based on qualitative review, validating the robustness of social media automated text processing.

Conclusions: This work demonstrates the value of social media data sources as a resource for in-depth studies of those affected by a dementia diagnosis and the potential to develop novel support systems based on their real time processing in line with the increasing digitalisation of medical care.
Introduction

The rise of dementia due to aging [1] is making an increasingly large contribution to disability-adjusted life years (DALYs) in society. There were an estimated 47 million people worldwide living with dementia in 2017 and this number is expected to reach 131.5 million by 2050 [2]. In the case of dementia, the diagnosis does not only impact upon the sufferer, owing to loss of autonomy, but also on caregivers and relatives, who may be risk of becoming overburdened [3]. The adjustments to lifestyle required to adapt to such a diagnosis are long-term and can fluctuate over time [4]. Therefore, it is crucially important to support sufferers and carers in these adjustments to improve their quality of life.

Internet usage and social media have been recognised as tremendous resources to support healthcare in decision making and to transform ongoing research [5-6]. In a recent study using Google Trends, Wang et al (2015) showed a plausible relationship between local internet search terms and new dementia cases and dementia-related outpatient visits in Taiwan [7]. Social media data can be used for digital disease detection, as demonstrated particularly for Twitter [8]. Such data has been used to develop classifiers to recognise depression in users [9], and to classify Twitter users who have attempted suicide from those who have not and from those that are clinically depressed [10]. Changes in mental health symptoms may also be detectable. For example, recent work by De Choudhury and colleagues has shown that by using social media data from Reddit [11], one can predict the likelihood of individuals shifting from posting about depression to posting about suicidal ideation [12].

These successful applications suggest that social media could also be helpful with identifying the familial impact and consequences of dementia diagnosis. In this study, we assess whether social media data gathered from Reddit can help improve our understanding of the effects of dementia diagnosis beyond targeted survey questions. The data obtained from Reddit are the unfiltered thoughts of people facing a dementia diagnosis and describe their personal circumstances in their own words. We used a previously established text analysis pipeline [13] and the open source tool carrot2 [14] to assess content and structure of posts revealing a dementia diagnosis.
Methods

Input data

Reddit is a social media platform that can be visited by registered and un-registered users. As of March 2018, Reddit reports 330 million registered users [15] that can exchange through topic-specific fora, so called subreddits. For example, Reddit hosts a subreddit called dementia [16], which describes itself as a community that supports individuals and families that need help when facing the consequences of a diagnosis or simply have questions about this medical condition. Once a user initiates communication through a post, others as well as the original poster can comment on the post, e.g. to answer questions. To analyse Reddit data relevant to dementia diagnosis, we downloaded user-collected datasets relating to posts and comments [16]. The data contained in both collections were gathered through the Reddit Application Programming Interface (API) [17].

As the focus here is on the impact of a dementia diagnosis on sufferers or caregivers, we only used the data set containing the posts in this study. The dataset comprised posts made between 01.01.2006 and 31.08.2015. As mentioned earlier, Reddit contains a multitude of subreddits and it is the user's choice which of the user-communities seem most suitable for their post. While there are a few dementia-specific subreddits, e.g. dementia, Alzheimers [18] or AlzheimersCanada [19], the number of posts available in these subreddits are small (low thousands). Furthermore, some of the consequences of a dementia diagnosis are potentially more relevant to other subreddits, which is why we conducted a keyword search as opposed to a subreddit search. We included all posts that either mention the word 'dementia' or 'Alzheimer' in their text and merged the results into one set. Removing posts that only contained a link to a web page led to a starting data set of 11,572 posts.

Selection of posts and manual curation

To derive a final data set, we applied a semi-automated approach, which means that we first automatically identified potentially relevant posts that were then manually curated. As suggested in earlier work by Coppersmith et al. [8], we automatically extracted posts that contained words indicating a diagnosis such as 'suffer' or 'diagnose', reducing the data set to 2,101 posts.

We note here that this process relies on the following assumptions: (i) the diagnosis has been established before and cited correctly by the poster and (ii) the post is not of a fictitious nature. We note further that a post can also be cross-posted (i.e. it may appear on two or more subreddits simultaneously with none of its contents changed).

We manually inspected the first 1,000 of these 2,101 posts (unique cross-posts removed) and manually annotated them with the diagnosis and the sufferer that had been reported. For this purpose, both the text and title of the post were taken into consideration. We note here that a single post may contain multiple diagnoses such as 'my grandmother and great-grandmother suffered from Alzheimer's'. Reported diagnoses could fall into one of three groups: (i) Alzheimer's, (ii) dementia and (iii) Alzheimer's, dementia. The latter group was assigned to posts that either mentioned two or more diagnoses, or there were
multiple people affected with several diagnoses. Occasional reports of non-Alzheimer
dementias, such as Frontotemporal dementia, Pick’s disease or Lewy body dementia, were
assigned a dementia diagnosis.

The manual curation was conducted by the first author of this paper (AO). The most
important relevance criteria for inclusion into the final data set was that a post contained
a factual statement such as ‘my grandmother suffers from dementia’. If the author of the
post expressed uncertainty or speculation, e.g. ‘we suspect she has dementia’, then this
post was excluded from the final data set. Following these procedures, we obtained a final
data set of 535 posts that were further processed and analysed.

Automated processing of final Reddit data

In a previous study on the classification of Reddit posts according to different mental
disorders [13], Gkotsis et al. developed a text processing pipeline that, among other
features, determines linguistic characteristics of Reddit posts. These linguistic
characteristics cover not only the length of the post in terms of numbers of words or
sentences, but also include measures of the complexity of a sentence, e.g. how many noun
phrases are contained in a sentence. We applied the same pipeline here to analyse the
Reddit posts contained in the final data set.

To further characterise the users in our data set and understand who is seeking support
online, we automatically predicted age and gender from user-related posts. For this
purpose, we used an existing, lexicon-based, linear regression prediction algorithm [20]
that uses the post’s text passages for both age and gender estimates. On Facebook
messages the algorithm was reported to possess an accuracy of 91.9% for gender
prediction and a Pearson correlation coefficient of 0.831 for age prediction (ages 13 to 60
separated into 12 bins). It was also noted that the algorithm works better with longer
texts, e.g. through concatenating individual messages. Therefore, if a person authored
more than one post in our data set (based on author name assigned to the post), all text
contents of the posts were merged together into one overall text and provided as input to
the algorithm. In all other cases, the prediction solely relies on the content of a single post.

Finally, we generated an Extensible Markup Language (XML) file from the posts and their
titles and supplied this to the existing open source tool carrot2 as an input file. This tool
automatically generates clusters from the textual content and auto-generates names for
each of these clusters. These cluster names provide an insight to, and summary of, the
content inside the posts. We used carrot2 with the default parameters to generate content
summaries of the posts. All source code for processing and handling the Reddit data that
are not part of the text processing pipeline, are available at
https://github.com/AnikaO/DementiaReddit.

Ethics statement

The King’s College London Ethics Committee granted exemption for this study. The data
were not gathered through interaction with any individuals nor is it identifiable private
information. Any extracts used for exemplary purposes were carefully paraphrased to
protect the privacy of individuals.
Results

Linguistic features of Reddit posts concerning dementia diagnoses

Posts containing details of a dementia diagnosis vary considerably in length, ranging from one to a maximum of 416 sentences in our data set. The mean number of sentences in a post is 33.8 sentences, with a standard deviation of 40.1. This suggests that in addition to the diagnosis, circumstantial information is recorded in most of the posts. The number of pronouns (both first and third person) used in the post also varies significantly from one to 1,143 pronouns per post, with a mean value of 77.6 pronouns per post (94.0 standard deviation).

Furthermore, we looked at the complexity of sentences by assessing the height of a sentence transformed into a tree structure based on the function of individual words in the sentence [21]. The greater the height of a sentence tree, the more complex a sentence is. In our data set, the sentence tree height ranged from four to 18 levels, with an average of 9.1 levels. This means that most sentences used to describe the personal circumstances are of a complex nature and not just short factual statements.

We investigated Reddit posts containing a dementia diagnosis statement according to five different criteria: (i) age and gender of the person posting, (ii) affected person and diagnosis, (iii) relevant subreddits authors are posting to, (iv) the types of messages posted and (v) the content of these posts overall.

(i) Age groups and gender of Reddit users posting dementia diagnoses

Figure 1 indicates the age groups of users, summarised by decade. The majority of post authors communicating about dementia diagnosis were aged between 20 and 40 years, with an average age of 30.1 years (dashed red line). Due to the way that age is predicted (lexicon-based linear regression), some outliers were expected which are illustrated as ages below ten or over 80 years. 56.1% of post authors are female according to the predictive algorithm. While both genders author posts, the majority of posts were predicted to be written by women.
To understand more about the dementia sufferers described in the post, we manually curated the affected person and the specific diagnosis from the content (text and title) of each of the posts. As illustrated in figure 2, the 10 people most often reported to suffer a type of dementia in Reddit posts are closely related to the writer, by means of a familial connection. Most posts relate to either the parents or grandparents of the poster. Interestingly, the number of times females (grandmother, mother, mother-in-law) are mentioned as sufferers is higher than their male counterparts. In our 535 posts comprising the final data set, only seven (1.3%) indicate that the writer themselves is suffering from dementia.

Figure 2 - A) Affected person and B) Diagnosis.
The most frequently disclosed diagnostic term mentioned in the posts was dementia with (278 posts; 52.0%), closely followed by Alzheimer’s (227 posts; 42.4%). In 5.6% (30) of the posts, either both types of diagnoses have been mentioned in conjunction with one person or there are several relatives suffering from different conditions. However, we note here that in some cases it is not possible to distinguish whether the author was intending to assign two diagnoses or whether one was used as clarification or as a synonym of the other.

(iii) Relevant subreddits authors are posting to

Figure 3 shows the top 10 most relevant subreddits identified by keyword search. The highest ranking subreddit is AskReddit, which is a generic user community open to any question a user may want to ask. In addition to the topic-specific communities (dementia and Alzheimer), there are other subreddits frequented such as depression or SuicideWatch. Furthermore, subreddits asking for specific advice, like Assistance, relationships and legaladvice, are represented. While nosleep might intuitively suggest the relevance to a symptom of dementia, this subreddit in fact consists of stories intended to frighten/scare the reader. As the stories can be of either fictional or non-fictional character, the posts made to this subreddit were still kept as part of the data set.

Figure 3 – Ten most common subreddits to reveal dementia diagnosis in posts
(iv) Types of Reddit posts

While manually curating dementia-related posts according to relationship to the affected person and the diagnosis, two major groups of posts became apparent: (i) people seeking to share their story and (ii) people requesting something from others. The latter group can be further divided into requests for advice or prayers, and “goods”. Interestingly, there is a group of posters on Reddit that are already coping with the situation and want to do something potentially beneficial for the diagnosed relative, e.g. producing a video or reprinting photographs. In these cases, the advice sought relates to how this can be accomplished. Other authors request the prayers of readers. “Goods” were of a material nature, e.g. people asking for postcards for the diseased to feel loved, or money. While some of the requests for money were on a charitable basis, i.e. people running marathons in memory of a relative dying of dementia, others used this to explain their predicament and asked for money directly to support medical bills.

(v) Content of Reddit posts

Using carrot2 in our final data set we obtained 66 groups of dementia-related posts with a topic assigned, and a group of 64 posts that did not fall into any of the 66 recognised topics. Not all of these groupings are necessarily meaningful in the context of a dementia diagnosis, but there are certainly groups indicating the effects and sufferers. Topics that indicate potential consequences of a diagnosis are for example ‘little money’, ‘caring for my mother’ and ‘didn’t remember’. We note here that there is also one collection of posts summarised as ‘feeling happy’. However, this does not necessarily indicate a positive connotation of “feeling happy”. For example, one of the posts falling into this group expresses that the author never feels happy. A positive expression of happiness alludes to gratitude, albeit the reason may be interpreted differently to the authors intention. For example, there are posts expressing gratitude over cannabis smoking facilitating the expression of suppressed feelings in relation to a dementia diagnosis. Groups reflecting the sufferer included ‘Grandmother is suffering from Alzheimer’ and ‘Mother was recently diagnosed’.

A complete list of all 66 topics is provided in supplementary file S1.
Discussion

In our study of Reddit posts the two most striking linguistic features were the mean number of pronouns per post and the sentence tree height. The high number of pronouns suggests that the posts in our data set contain personalised /writer-relevant information and the high mean sentence tree height indicates the complexity of the sentences. We found that the authors of posts revealing a dementia diagnosis were mostly women, predominantly aged 20 to 40 years, writing about close relatives (parents and grandparents), and primarily about female relatives.

While the algorithm we used to automatically determine the age and gender of authors was developed using Facebook data [20], it was further evaluated on blogs and tweets. We judged this tool to be suitable to predict age and gender of the authors in our Reddit post dataset. However, it has been recognised in the past that using textual content from online media to predict age can be insufficient [22]. We plan to conduct an additional evaluation on our dataset once it contains significantly more posts in which age and gender could be manually curated from the data based on a mention in the title or text of the post.

In a report issued by the Pew Research Center's Internet & American Life project [23], it was reported that young males (18-29 years) are particularly likely to use Reddit. The same report also showed that only 5% of female internet users aged 18 to 49 are using Reddit. Our finding of mostly females authoring Reddit posts concerning a dementia diagnosis is therefore particularly relevant, as it diverges from the most prevalent user group. The age groups we identified as relevant are in line with the most prevalent age groups for female users given in the report. This suggests that those users communicating on Reddit about dementia diagnosis are not the most represented user group on this social media site.

Our results from the manual curation of the diagnosis suggests that there is a small proportion of posts (<5.6%) that mention both a diagnosis for ‘Alzheimer’s and dementia’ simultaneously. This may be related to earlier findings that the sufferer themselves is unsure about the diagnosis, not just because of the symptoms of the condition. It has been found in earlier work that only 47% of dementia sufferers know their correct diagnosis [24]. As we are relying on authors of posts to correctly report the diagnosis and this information being provided based on a statement of fact, it is impossible for us to resolve uncertain cases. Therefore keeping a group with both diagnoses together is an acceptable option.

Our results also demonstrate that authors of posts containing dementia diagnosis are seeking support outside topic-specific communities. While posts concerning a dementia diagnosis appear in topic-specific communities (e.g. dementia), other communities that represent the pragmatic reality of a dementia diagnosis (e.g. legaladvice for questions relating to a Power of Attorney) were represented. The content analysis revealed pragmatism in relation to a dementia diagnosis with people asking for legal or caregiver advice.

The data studied in traditional survey-based questionnaires are answers to predefined questions. Here, our input data goes beyond interview questions to get an "unfiltered perspective" of the major concerns of relatives and sufferers with respect to a dementia
diagnosis. It has been shown in the past that a dementia diagnosis can lead to depression and anxiety in relatives, especially those that are primary caregivers [25] and some of the subreddits (depression, SuicideWatch, offmychest) identified as prevalent in our data set also allude to this fact. In response programmes combining face-to-face coaching with tailored Web-based modules to reduce overburdening of caregivers are currently being trialled [26].

In addition, we found that there were a number of posts made to the subreddit nosleep, which aims at unsettling the reader with (non-)fictional stories. The presence of dementia diagnosis in these stories suggests that there are aspects related to dementia that can instil fear in others, not necessarily facing a diagnosis themselves and therefore potentially increase stigma towards this condition [2].

While social media offers an additional data source to study the impact of dementia diagnoses, its processing could also develop into a means of offering support to caregivers through appropriate, targeted interventions. Although such interventions would not currently be suitable for sufferers due to the low number of posters writing about their own condition, this is likely to change over time as younger generations engage with social media more, and healthcare becomes increasingly digitalised.

Conclusion

In our study we demonstrated that posts made on the social media platform Reddit can be used to study user characteristics and better understand familial consequences of a dementia diagnosis. Mainly female authors write posts about dementia diagnoses to user communities they can seek advice and support from. Most of the posts in our data set are written about a close relative to the author, predominantly of a female gender. In addition to people just sharing their personal circumstances, others do ask for support in various ways in topic-specific and general communities.

Despite our results being based on a small data set, the trends that have been identified here are consistent with those previously reported based on surveys and qualitative reviews. This suggests that social media can already be used for studying more in-depth responses of carers and relatives of dementia sufferers, going beyond a questionnaire structure that can be restrictive or even bias the answers of the respondent. Due to the interactive nature of Reddit, this social media source could be used for targeted interventions to support caregivers and relatives of someone with dementia.
Acknowledgments

This paper represents independent research part-funded by the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King’s College London. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

RJBD's work is also supported by the NIHR University College London Hospitals Biomedical Research Centre, and by awards establishing the Farr Institute of Health Informatics Research at UCLPartners, from the Medical Research Council, Arthritis Research UK, British Heart Foundation, Cancer Research UK, Chief Scientist Office, Economic and Social Research Council, Engineering and Physical Sciences Research Council, National Institute for Health Research, National Institute for Social Care and Health Research, and Wellcome Trust (grant MR/K006584/1).

TJPH would like to acknowledge King’s College London and the NIHR Biomedical Research Centre at Guy’s and St Thomas’ NHS Foundation Trust and the NIHR Biomedical Research Centre for Mental Health.

RD is supported by a Clinician Scientist Fellowship from the Health Foundation in partnership with the Academy of Medical Sciences (https://www.acmedsci.ac.uk/grants-and-schemes/grant-schemes/csf/) and RD and GG are funded by the e-HOST-IT research programme.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.
References


Data analysed: user-collected data set from https://redd.it/3mg812 for posts and https://redd.it/3bxlg7 for comments.


Supporting File S1 - S1_file.csv

Automatically generated groups of posts. This file contains a list of 66 generated named groups of posts obtained by applying Carrot2. For each of the groups, the file indicates how many of the 535 posts fall within each particular group.

The data used for the analysis in this study are posts available through the Reddit API (https://www.reddit.com/dev/api), which requires registration and authentication for download.