DOI:
10.1159/000509123

Research Article

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

Author's accepted manuscript. DOI: https://doi.org/10.1159/000509123

George Gkotsis[1], Christoph Mueller[1,2], Richard J.B. Dobson[3,4], Tim J.P. Hubbard[5], Rina Dutta*[1,2]

[1] Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, SE5 8AF, UK

[2] South London and Maudsley NHS Foundation Trust, London, UK

[3] Medical Research Council (MRC) Social, Genetic & Developmental Psychiatry Centre (SGDP), King's College London, London, SE5 8AF, UK

[4] Farr Institute of Health Informatics Research, UCL Institute of Health Informatics, University College London, London, WC1E 6BT, UK

[5] Department of Medical & Molecular Genetics, King's College London, London, SE1 9RT, UK

Short Title: Dementia Diagnosis and Social Media

*Corresponding Author

Rina Dutta, King's College London, United Kingdom and South London and Maudsley NHS Foundation Trust, London, UK
rina.dutta@kcl.ac.uk

Senior Clinical Lecturer / Consultant Psychiatrist
Department of Psychological Medicine, Division of Academic Psychiatry
PO Box 84 | Room E3.07 | 3rd Floor East Wing | IoPPN
King's College London | De Crespigny Park | London SE5 8AF
Tel +44 (0)7904 207378 | Fax +44 (0)207 848 5408

Number of Tables: none (1 supplementary table)

Number of Figures: 3 figures (2 supplementary figures)

Word count: 2,803

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

Abstract

Introduction: Caregivers for people with dementia face a number of challenges as changing family relationships, social isolation or financial difficulties. Internet usage and social media are increasingly being recognised as resources to increase support and general public health.

Objective: The aim of this study was to explore using automated analysis (i) the age and sex of people who post to the social media forum Reddit about dementia diagnoses, (ii) the affected person and their diagnosis, (iii) which 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 annotation of 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 / clustering 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 females, 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 person's diagnosis but also potential outcomes, for example hardships experienced by the caregiver or the requirement for legal support.

Conclusions: This work demonstrates the value of social media data sources as a resource for in-depth studies of those caregivers affected by a dementia diagnosis. It is important to study those groups actively posting online, both in topic-specific and general communities, because they are the ones most likely to benefit from novel internet-based support systems or interventions.

Introduction

Dementia is considered one of the major public health challenges of the 21st century making an increasingly large contribution to disability-adjusted life years in society [1]. However, the diagnosis does not only impact upon the person with dementia, but also on relatives and caregivers, who are at risk of becoming overburdened [2, 3]. The adjustments in lifestyle required to adapt a neurodegenerative condition need to be long-term and can fluctuate over time [4]. Therefore, it is crucially important to support persons with dementia and their caregivers in these adjustments to improve their quality of life.

Internet usage and social media have been recognised as emerging resources to support healthcare in augmenting existing and creating new public-health capabilities [5]. 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 [6]. Social media data can be used for digital disease detection, as demonstrated particularly for Twitter [7]. 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 (https://www.reddit.com/), one can predict the likelihood of individuals shifting from posting about depression to posting about suicidal ideation [8].

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 caregivers and relatives of people facing a dementia diagnosis and describe their personal circumstances in their own words. We used a previously established text analysis pipeline [9] and the open source text clustering tool carrot2 [10] to assess content and structure of posts revealing a dementia diagnosis.

Materials and Methods

Input data

Reddit is a social media platform that can be accessed by registered and un-registered users. As of March 2018, Reddit reports 330 million registered users that can exchange through topic-specific fora, so called subreddits. For example, Reddit hosts a subreddit called r/dementia [11], 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. The data contained in both collections were gathered through the Reddit Application Programming Interface (API) [12].

As the focus here is on the impact of a dementia diagnosis on people with dementia or caregivers, we only used the dataset containing the original posts (excluding the comments) for this study. The dataset comprised posts made between 1st January 2006 and 31st August 2015. Other mental health research studies based on Reddit data have focussed on specific diagnosis-related subreddits [13]. However, it is the user's choice which of the user-communities they deem most suitable for their post and there are only a few dementia-specific subreddits, (e.g. r/dementia, r/Alzheimers or r/AlzheimersCanada), with the number of posts available in these subreddits being relatively small (in the low thousands). Furthermore, some of the consequences of a dementia diagnosis are potentially more relevant to other subreddits, which is why we chose the distinct methodology of conducting 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.

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

Selection of posts and manual annotation

Removing posts that only contained a link to a web page led to a starting data set of 11,572 posts. 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 annotated. As suggested in earlier work by Coppersmith et al. [7], we automatically extracted posts that contained words indicating a diagnosis such as ‘suffer’ or ‘diagnose’, reducing the data set to 2,101 posts.

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 person with dementia 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 diagnostic statements 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 vascular dementia, frontotemporal dementia, Pick's disease or Lewy body dementia, were assigned a dementia diagnosis.

The most important relevant criterion for inclusion in the final data set was that a post contained a factual statement such as 'my grandmother has 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 [9], Gkotsis and colleagues 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 predicted age and sex from user-related posts. For this purpose, we used an existing weighted lexicon [14] that uses the post's text passages to predict both age and sex. The lexicon (words and weights) were trained from word usage in Facebook, Blog and Twitter data with associated demographic labels.

Finally, we generated an Extensible Markup Language (XML) file from all the posts and their titles and supplied this to the existing open source tool carrot2 [10] as an input file (https://doc.carrot2.org/). This tool supports multiple algorithms for document clustering and generates these clusters from the textual content assigning each with a label. Each cluster has a reference to all documents that

belong in them. These cluster names therefore provide an insight to, and summary of, the content inside the posts, allowing us to gain an overview of the topics described in our documents, how the documents are distributed across these topics, as well as the affinity between the topics themselves. 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.

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 widely from one to 1,143 pronouns per post, with a mean value of 77.6 pronouns per post (94.0 standard deviation). This is considerably higher than the mean value of 56.5 pronouns per post (86.7 standard deviation) found in all Reddit posts containing the keywords dementia or Alzheimer.

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 [15]. 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 (see Supplementary Figure 1A and 1B for synthetic examples of sentences trees).

Age groups and sex 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 female/male authors post, the majority of posts were predicted to be written by females.

Affected person and diagnosis

To understand more about the dementia described in the post, we manually annotated the affected person and the specific diagnosis from the content (text and title) of each of the posts. As illustrated in Figure 2, Reddit post about dementia most often discuss those who are closely related to the writer, often by means of familial connection, such as parents or grandparents. Females were

mentioned more frequently as person with dementia than males. In our 535 posts comprising the final data set, only seven (1.3%) indicate that the writer themselves has dementia.

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 dementia and Alzheimer's disease were mentioned in conjunction with one person.

Which subreddits authors are posting to?

Figure 3 shows the top 10 subreddits with the most frequent mention of the keywords 'dementia' or 'Alzheimer' in their text. The highest ranking subreddit is r/AskReddit, which is open to any question a user may want to ask. In addition to the topic-specific communities (r/dementia and r/Alzheimer), there are other subreddits frequented such as r/depression or r/SuicideWatch. Furthermore, subreddits asking for specific advice, like r/Assistance, r/relationships and r/legaladvice, are represented. While r/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.

Types of Reddit posts

While manually annotating 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 photograph. In these cases, the advice sought relates to how this can be accomplished. "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 asked for money to support medical bills directly, using their posts to explain their predicament and financial hardship.

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 on people with dementia and their carers. 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 author’s intention. For example, there are posts expressing gratitude over cannabis smoking facilitating the expression of suppressed feelings in relation to a dementia diagnosis. A complete list of all 66 topics is provided in Supplementary Table 1.

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. Posts mentioning dementia diagnoses, compared to all Reddit posts containing the keywords 'dementia' or 'Alzheimer' had comparatively more pronouns indicating higher inter-personal focus.

We also found that the authors of posts revealing a dementia diagnosis were mostly female, predominantly aged 20 to 40 years, writing about close relatives (parents and grandparents), and primarily about female relatives. This reflects research in the medical literature, which shows higher prevalence, more disability and longer survival amongst females [16], as well as higher levels of female caregiving internationally [17].

The predicted gender skew of the majority (56.1%) of post authors being female is contrary to most subreddits which are posted to by males, including the most commonly used front page r/AskReddit (as shown by Burkhart using data through the end of November 2017; http://bburky.com/subredditgenderratios/). In this analysis there were only two subreddits with predominantly female posters: r/relationships (51.6%) and r/TwoXChromosomes (62.4%).

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. r/dementia), other communities that represent the pragmatic reality of a dementia diagnosis (e.g. r/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 persons with dementia 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 [18] and some of the subreddits (r/depression, r/SuicideWatch, r/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 [19].

In addition, we found that there were a number of posts made to the subreddit r/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.

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 persons with dementia due to the low number of posters writing about their own condition, this is likely to change over time as successive generations will have used social media since childhood as part of everyday life, and as healthcare becomes increasingly digitalised.

Limitations

While the algorithm we used to automatically determine the age and sex of authors was developed using Facebook data [14], it was further evaluated on blogs and tweets. We judged this tool to be suitable to predict age and sex 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 [20].

There may be selection bias inherent in using Reddit to study dementia-related topics. Although it is a social media platform widely used amongst a broad and diverse population, with less expectation of privacy than on other channels, it may not represent those relatives / care-givers suffering consequences of the diagnoses. Such individuals may seek deeper advice or support on carer or legal issues, for example, using specialist fora such as https://www.dementiacarecentral.com/forum/ or https://www.dementiaforum.org/. Informed consent to study posts in these fora would be helpful for a richer understanding of the impact of dementia diagnosis on caregivers and relatives, which could help further in developing earlier interventions.

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 female relative

of the author. 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 persons with dementia, 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.

Statements

Acknowledgement

The authors would principally like to acknowledge the contribution of Dr Anika Oellrich, who had the initial idea for the study, and was supported by GG and RD in its development and also helped to analyse the data.

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 and Social Care.

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.

Statement of Ethics

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.

Conflict of Interest Statement

The authors have no conflicts of interest to declare.

Funding Sources

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. CM receives salary support from 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 supported by the NIHR University College London Hospitals Biomedical Research Centre, and by awards establishing the Farr Institute of Health Informatics Research at UCL Partners,

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

Author Contributions

GG and RD conceived of the study, with the initial idea being Dr Anika Oellrich's; GG and RD participated in its design and wrote the manuscript. GG and RD analysed the data. CM, RJBD and TJPH helped to draft the manuscript. All authors read and approved the final manuscript.

References:

1. Murray CJ, Vos T, Lozano R, Naghavi M, Flaxman AD, Michaud C, et al. Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990-2010: a systematic analysis for the Global Burden of Disease Study 2010. Lancet. 2012 Dec 15;380(9859):2197-223.
2. Derksen E, Vernooij-Dassen M, Gillissen F, Olde Rikkert M, Scheltens P. Impact of diagnostic disclosure in dementia on patients and carers: qualitative case series analysis. Aging Ment Health. 2006 Sep;10(5):525-31.
3. Bjoerke-Bertheussen J, Ehrt U, Rongve A, Ballard C, Aarsland D. Neuropsychiatric symptoms in mild dementia with lewy bodies and Alzheimer's disease. Dement Geriatr Cogn Disord. 2012;34(1):1-6.
4. Laakkonen ML, Raivio MM, Eloniemi-Sulkava U, Tilvis RS, Pitkala KH, Pitkala KH. Disclosure of dementia diagnosis and the need for advance care planning in individuals with Alzheimer's disease. J Am Geriatr Soc. 2008 Nov;56(11):2156-7.
5. Dredze M. How Social Media Will Change Public Health. IEEE Intelligent Systems. 2012;27(4):81-84.
6. Wang HW, Chen DR, Yu HW, Chen YM. Forecasting the Incidence of Dementia and Dementia-Related Outpatient Visits With Google Trends: Evidence From Taiwan. J Med Internet Res. 2015 Nov 19;17(11):e264.
7. Coppersmith G, Dredze M, Harman C. Quantifying Mental Health Signals in Twitter. Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality. Baltimore, Maryland, USA2014. p. 51-60.
8. De Choudhury M, Kiciman E, Dredze M, Coppersmith G, Kumar M. Discovering Shifts to Suicidal Ideation from Mental Health Content in Social Media. Proc SIGCHI Conf Hum Factor Comput Syst. 2016 May;2016:2098-110.
9. Gkotsis G, Oellrich A, Hubbard T, Dobson RJB, Liakata M, Velupillai S, et al. The language of mental health problems in social media. The Third Computational Linguistics and Clinical Psychology Workshop (CLPsych). 2016. p. 63-73.
10. Weiss D, Osinski S. Carrot2 software.
11. Reddit. Data analysed: user-collected data set from https://redd.it/3mg812 for posts and https://redd.it/3bxlg7 for comments.
12. Reddit. Reddit API documentation. 2018.
13. Gkotsis G, Oellrich A, Velupillai S, Liakata M, Hubbard TJ, Dobson RJ, et al. Characterisation of mental health conditions in social media using Informed Deep Learning. Sci Rep. 2017 Mar 22;7:45141.
14. Sap M, Park G, Eichstaedt J, Kern M, Stillwell D, Kosinski M, et al. Developing Age and Gender Predictive Lexica over Social Media. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar: Association for Computational Linguistics; 2014. p. 1146-51.
15. Charniak E. Statistical Techniques for Natural Language Parsing. AI Magazine. 1997 12/15;18(4).
16. Sinforiani E, Citterio A, Zucchella C, Bono G, Corbetta S, Merlo P, et al. Impact of gender differences on the outcome of Alzheimer's disease. Dement Geriatr Cogn Disord. 2010;30(2):147-54.
17. Joling KJ, Windle G, Droes RM, Meiland F, van Hout HP, MacNeil Vroomen J, et al. Factors of Resilience in Informal Caregivers of People with Dementia from Integrative International Data Analysis. Dement Geriatr Cogn Disord. 2016;42(3-4):198-214.
18. Joling KJ, van Hout HP, Schellevis FG, van der Horst HE, Scheltens P, Knol DL, et al. Incidence of depression and anxiety in the spouses of patients with dementia: a naturalistic cohort

study of recorded morbidity with a 6-year follow-up. Am J Geriatr Psychiatry. 2010 Feb;18(2):146-53.
19. Boots LM, de Vugt ME, Smeets CM, Kempen GI, Verhey FR. Implementation of the Blended Care Self-Management Program for Caregivers of People With Early-Stage Dementia (Partner in Balance): Process Evaluation of a Randomized Controlled Trial. J Med Internet Res. 2017 Dec 19;19(12):e423.
20. Nguyen D-P, Trieschnigg RB, Dogruoz AS, Gravel R, Theune M, Meder T, et al. Why Gender and Age Prediction from Tweets is Hard: Lessons from a Crowdsourcing Experiment. Proceedings of the 25th International Conference on Computational Linguistics, COLING 2014. 2014 2014/08/23/:1950-61.

Figures:

Figure 1: A) Age and B) Sex distribution of Reddit post authors

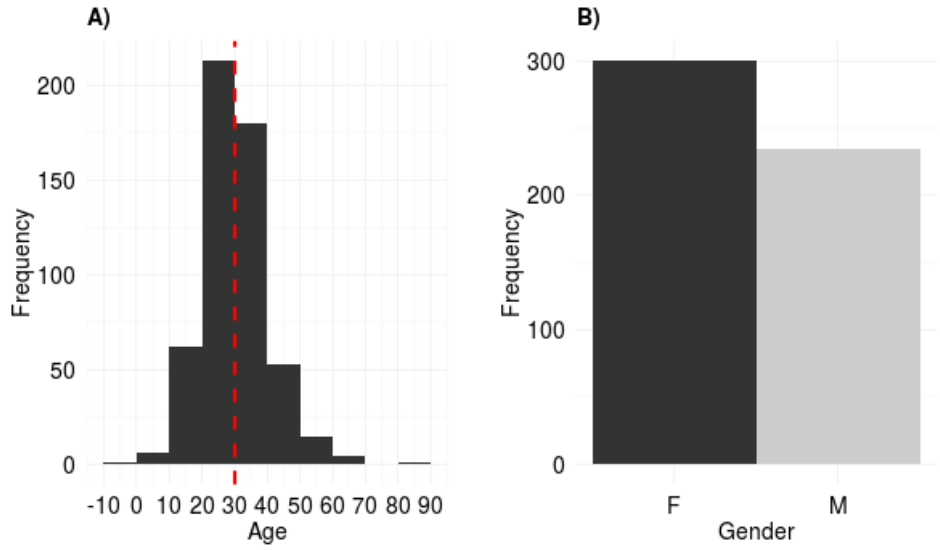


Figure 2: A) Affected person and B) Diagnosis.

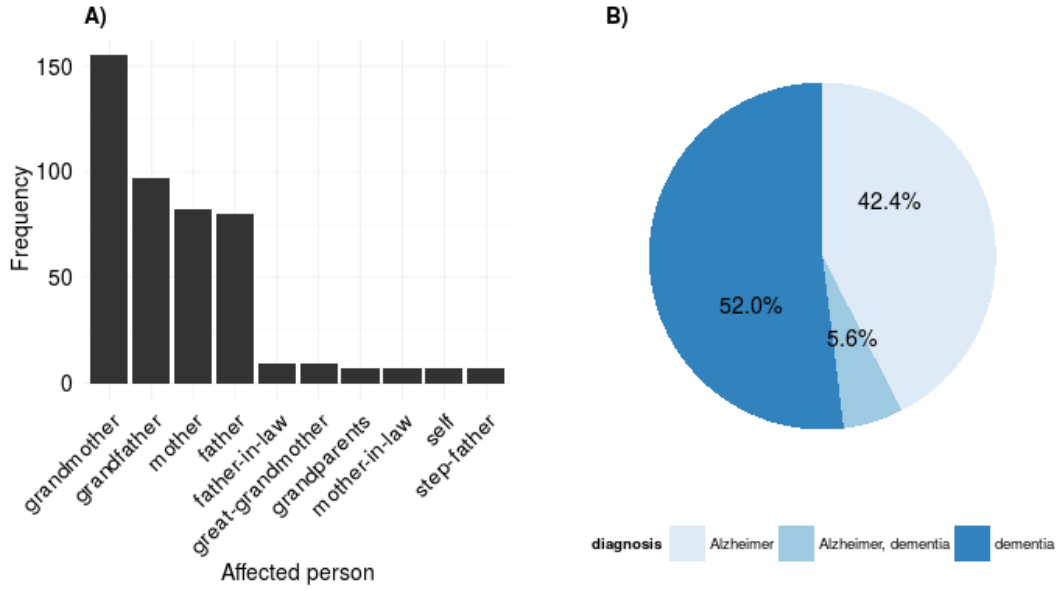
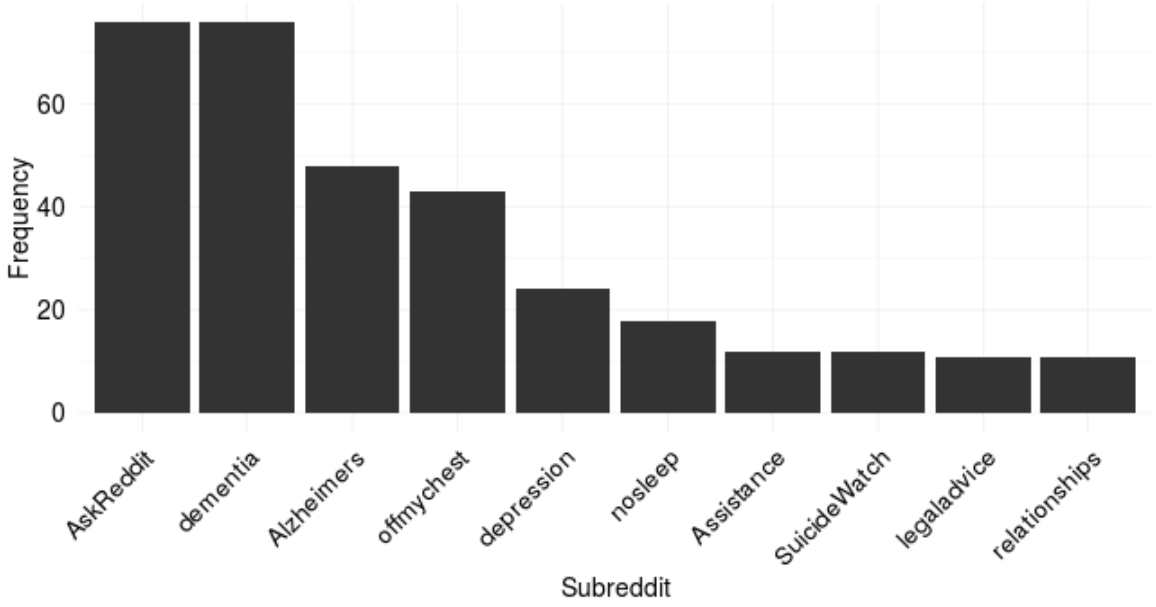


Figure 3: Ten most frequent subreddits to reveal dementia diagnosis in posts


Frequency
60
40
20
0
AskReddit
dementia
Alzheimers
offmychest
depression
nosleep
Assistance
SuicideWatch
legaladvice
relationships
Subreddit

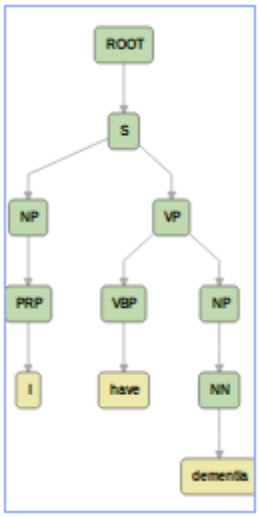
Supplementary Tables & Figure

Supplementary Table 1: Post groups with numbers of posts

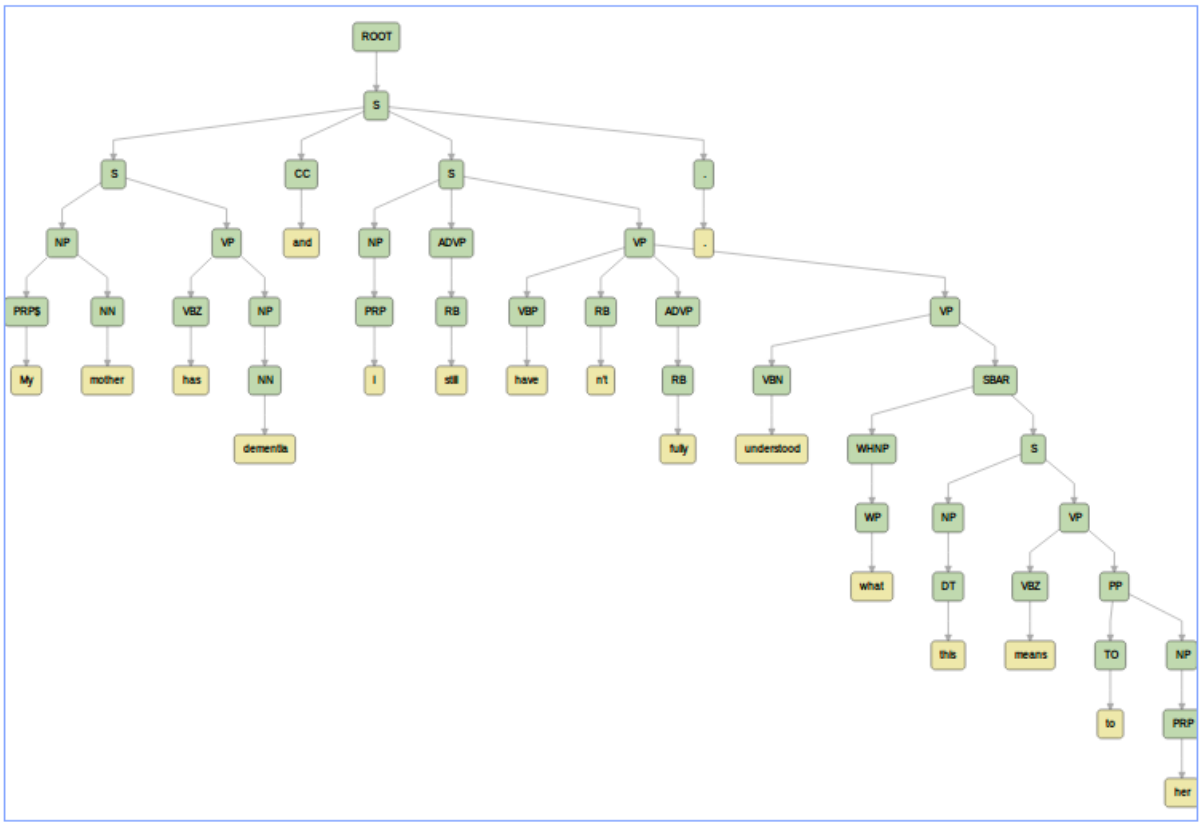
Name of post group	Number of posts
"Caring for my Mother	127
"Mom and Dad"	93
"Father Died"	88
"Died of Alzheimers"	87
"Loving Person"	79
"Took a Look"	72
"Willing to Read"	71
"Love my Parents"	67
"Past Week"	65
"Stopped Working"	65
"Able to Work"	63
"Dad Took"	62
"Stopped Coming"	62
"Little Old"	61
"Taking her Medications"	61
"Father Passed"	59
"Told this Story"	59
"Great Start"	58
"Long as I can Remember"	58
"Able to Tell"	57
"Close Friends"	57
"Mother's Sister"	57
"Feeling a Bit"	56
"Felt Really"	56
"Mom Lost"	56
"Able to Talk"	55
"Entire Family"	55
"Feeling Happy"	55
"Mom Took"	55
"New Life"	55
"Possibly Help"	55
"Probably Know"	55
"Medical Help"	54
"Moved to a Better"	54
"Wrong Thing"	54
"Grandmother is Suffering from Alzheimers"	53
"Left that Job"	53
"Lives Far"	53
"People I Knew"	53
"Grandma has Dementia"	52
"Mother was Recently Diagnosed"	51
"New House"	51
"Couple Days"	50
"Didn't Leave"	50
"Little Money"	50

"Finally Called"	49
"Grandfather was Diagnosed with Alzheimers"	49
"Grandmother was Diagnosed with Dementia"	49
"Lot of Money"	49
"Mother is Actually"	49
"Parents Left"	49
"Didn't Remember"	48
"Nights Ago"	48
"Old Age"	47
"Little Sister"	43
"Little Brother"	41
"Caring for someone with this Disease"	38
"Walked out of the Room"	35
"Fucking Bad"	34
"Fucking Hard"	34
"Hospital Room"	34
"Great Aunt"	19
"Great Uncle"	18
"Theirs and were a Joy"	11
"Lasting Power of Attorney"	10
"Crazy Uncle"	9

Supplementary Figure 1A: Synthetic example of a short sentence tree: “I have dementia”



Supplementary Figure 1B: Synthetic example of a taller sentence tree: “My mother has dementia and I still haven’t fully understood what this means to her.”



(These are synthetic examples, not taken from the text due to licensing).