ENCEPHALOS 57, 1-12, 2020

Analysis of Spontaneous Speech Using Natural Language Processing Techniques to Identify Stroke Symptoms

E. S. RIGAS[1], M. PAPOUTSOGLOU[1], K. TSAKPOUNIDOU[2], A. SERDARI[3], E. MOUZA[4], H. PROIOS[2,4]

Abstract

This study measures the frequency of detections for Face, Arms, Speech symptoms, using natural language processing techniques, after extrapolating them from medical history intakes of stroke survivors. We use a dataset of 58 stroke survivors, who were undergoing diagnostic screening for communication impairments. All participants have responded orally, and audio recordings had been transcribed for subsequent analysis of the text. We used two natural language processing techniques, namely term frequency (tf) which shows the number of times a word occurs in a document and latent Dirichlet allocation (LDA) which distinguished 3 topics in text analysis as the ideal number of topics. Regarding tf, we rank single, double and triple word sequences. For single words, we observed that "speech" and "arm" appeared most frequently in the medical history intake (frq~ 16.1% and 6.9% respectively), whereas "face" and other related words were less frequent (frq~ 2.68%). Regarding LDA, 3 topics revealed that "arm" and "speech" were clearly distinguished from each other. "arm" and "suddenly" in Topic 1, "stroke", "reports" and "weakness" in Topic 2, "hospital" and "weakness" in Topic 3, respectively, were detected. Results show evidence that stroke symptoms should be included in clinical decision making for health care practice.

Key words: Stroke; Term frequency; Latent Dirichlet Allocation; Symptoms; Medical history

[1]Aristotle University of Thessaloniki, Department of Informatics, Thessaloniki, Greece.

[2]University of Macedonia, Department of Educational and Social Policy, Thessaloniki, Greece.

[3]Democritus University of Thrace, University Hospital of Alexandroupolis, Department of Child and Adolescent Psychiatry, Alexandroupolis, Greece.

[4]Department of Speech Therapy, Anagennisis (Revival) Physical Recovery and Rehabilitation Centre, Nea Raidestos, Filothei, Thessaloniki, Greece.

Corresponding author: Hariklia Proios

Department of Educational and Social Policy, University of Macedonia, Egnatia 156 Str., Thessaloniki, Greece GR-54636, hproios@uom.edu.gr, +30 2310 891630

1. Introduction

There are over 33 million stroke survivors according to the Global Burden of Disease Study 2010 (Feigin et al., 2014), of which over one million are affected with aphasia in the United States alone (Moffatt, Pourshahid, & Baecker, 2017). There have been several efforts from research towards identifying genetic biomarkers for stroke (Bang, 2017; Damico, Müller, & Ball, 2010; Katan, & Elkind, 2018; Kim, Moon, & Bang, 2013). This knowledge is immensely important for ensuring healthy living in a world where the prevalence of people with stroke related disability is increasing (Feigin, Norrving, & Mensah, 2017; Huang et al., 2016). In this vein, the history intake is an extremely valuable part of individual testing and treatment. During the patient medical history intake, a common question to a patient is "Why are you here?". In this study, we aim to analyze medical history intake in order to identify the face, arm, speech symptoms that are described by the patients who are stroke survivors.

The acronym FAST which represents face, arms, speech and time is used to identify the symptoms that indicate someone is having a stroke (Dombrowski et al., 2015) and time for ambulance help. The most common stroke symptoms are F: facial weakness, A: arm hemiplegia/hemiparesis or S: speech deficits and/or aphasia.

In this study, we create a model that links face, arm and speech symptoms to individual verbal productions. Often times, when taking a medical history intake of a stroke survivor, communication is impaired. The verbal output can often be incoherent and incohesive, or even unintelligible (Kurowski, & Blumstein, 2016). In this study, we used transcribed responses to a "Why are you here today?" question taken from the Greek version of the Aachen Aphasia Test (AAT) (Proios et al., 2006) to understand and identify the relevant descriptors of the three symptoms mentioned by the stroke survivors. We first examined the spontaneous dialogues and identified personal descriptors included stroke symptoms. Thus, we explored the relationship between pairs of symptoms and topics revealed from the modeling process.

To avoid diagnostic pitfalls of individualized reports the scope of the history intake was refined using computer based techniques. The main purpose of our study was the application of a standard natural language processing technique, identifying symptoms of transcribed answers.

Although there are studies on computer-based techniques intended for treatment or therapeutic purposes (Grawemeyer, Cox, & Lum, 2000; Vannobel, & Toulotte, 1992) as well as diagnosis of aphasia (Sponsler, & Burkhart, 2016), to the best of our knowledge there is limited evidence of stroke symptoms from medical history intake. There is some computer-based research on identifying the topics stroke survivors are interested in talking about (i.e. 'food and drink', 'nature and gardening', 'entertainment' and so on) (Palmer, Hughes, & Chater, 2017) and on assessing stroke survivors' needs for relevant information about stroke and its consequences (Dixon, Thornton, &Young, 2007; Mangset et al., 2008; Worrall et al., 2011).

2. Materials and methods

We use a dataset consisting of 58 participants, 68.96% male and 31.04% female, aged between 26 and 78 with mean 58.29 and median 61, the majority having suffered an ischemic stroke. All participants were asked the question "why are you here today?". Based on this, they described their symptoms orally which they have been typewritten by the authors. Each patient's story was summarized manually in order to keep the important information and reduce the noise, aiming to improve the efficiency of the analysis. The stories were written in the Greek language. Prior to the analysis, a pre-processing step took place. At this step, we defined a dictionary of stop words (e.g., and, his/her) in order to keep the set of words that provide a conceptual meaning. To do so, 1) the list of prepositions was added, 2) the list of connecting words was added and 3) other words that were found in our experiments not to give extra meaning were also added. Note that such digital dictionaries of stops words are commonplace in the English language, but none existed in Greek.

The pre-processing phase consists of the following steps:

1. Remove the names of patients from the raw text. This step took place also for anonymization manners.

2. Turn all text to lower case.

3 Remove all special characters such as "." or ",". To achieve this, we used the regular expression "[[:punct:]][$]", where punct is defined by the programming language R that was used in this work. In computer science, a regular expression is defined as a sequence of characters that define a search pattern. We selected to remove punctuation symbols such as comma, docs, semi columns and other symbols as they do not provide any meaningful information for our needs. Additionally, by removing punctuation we also remove illegal characters which could produce errors in the text formation.

4 Remove all stop words.

After the pre-processing phase, the text analysis took place. To analyze the text, we used two standard natural language processing techniques, namely term frequency (tf) (Salton, & Buckley, 1998) and latent Dirichlet allocation (LDA) (Blei, Ng, & Jordan, 2003).

The tf, which is part of the TFIDF, short for term frequency–inverse document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus (de Zubicaray et al., 2006). For the term frequency tf(t,d), the simplest choice is to use the raw count of a term in a document, i.e., the number of times that term t occurs in document d.

In natural language processing, latent Dirichlet allocation (LDA) is a generative statistical model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar. In LDA, each document may be viewed as a mixture of various topics where each document is considered to have a set of topics that are assigned to it via LDA. Figure 1 shows the procedure that follows to calculate the topics and assign the terms in the topics. α is the parameter of the Dirichlet prior on the per-document topic distributions, β is the parameter of the Dirichlet prior on the per-topic word distribution, θ_d is the topic distribution for documents, φ_k is the word distribution for topic k, z_{dj} is the topic for the j-th word in document d, and w_{dj} is the specific word. Figure 2 shows the whole text analysis procedure.

Aiming to implement the methodology steps described previously, we selected the open source option. Methodology steps contain qualitative analysis of raw digitalized content and need special handling in contrast to simpler approaches needed for quantitative data. Many different solutions exist for qualitative analysis of text data and some of them can save time in the hands-on process and help researchers who don't have coding experience. This solution comes from tools appropriate for qualitative analysis and one of the most frequently used in NVivo (https://www.qsrinternational.com/nvivo/nvivo-products) (Abbe et al., 2016; Roundtree, 2018; Ward, Gbadebo, & Baruah, 2015). Even though NVivo constitutes a user-friendly software for content analysis with an interface available for all level of users, its use still has many drawbacks. It is commercial software and demands time to learn how to use it as it is a semi-automated tool (Ranney et al., 2014). Aiming to avoid these disadvantages we selected to implement the process using the open software solution provided by the R programming language. Our approach is semi-automated only in the first step of the transformation of raw handwritten content into digital. After that, the content analysis steps were fully automated under an open-source solution.

3. Results

Aiming to extract meaningful results from the text, we selected to run the fourth step from fig. 2 (i.e., Analysis of text) for three different n-gram cases. From our raw text we collected a contiguous sequence of words of n items. For our case, the items are the sequence of words and we selected to run the text analysis step for three different cases. More specifically for unigram (n = 1), bigram (n = 2) and trigram (n = 3). We selected the n-gram option as in some cases we found meaningful results for n > 1.

Regarding tf we observe that "speech" and "arm" appear frequently, while words related to face, or other conditions also exist (see Tables 1-3). In more detail, for the single word frequency, the words *speech, weakness* and *felt* appear more often. For the double-word frequency, the pairs *inability talk, weakness arm* and *can not* appear more often. Finally, for the three-word frequency, the triples *can not talk*, *lose senses arm* and *started weaken finally* appear more often. Overall, someone can observe that the most frequent words refer to symptoms related to speech and arm, while the face also appears

but with smaller frequency.
Regarding LDA we have to mention that there is specific number of produced topics which is the most suitable for every situation according to the analysis. After experimental process we identified that the number of topics which separates and explains better our needs is the selection of 3 topics. So, we selected these 3 topics which are clearly distinguished from each other and we observe that "speech" and "arm" are categorized in two of them, while the third one is vaguer as it contains words related to the general condition of the patients (see Figures 3-5). In this way, the results of the tf are confirmed as these two symptoms are selected by

4. Discussion

In this study we try to investigate if there is a relationship between face, arm, speech stroke symptoms and the transcribed spontaneous verbal productions of stroke survivors. The tf and LDA processing techniques were applied to assess the participants' spontaneous verbalization. It has been shown that the words 'speech' and 'arm' are the most commonly mentioned. For the double-word frequency, the pairs 'inability talk', 'weakness arm' and 'can not' appear more often, while for the three-word frequency, the triples 'can not talk', 'lose senses arm' and 'started weaken finally' appear more often. Our findings are in line with other studies pointing that the face, arm, speech symptoms are identified by the vast majority of stroke survivors (Kleindorfer et al., 2007). Other stroke symptoms are mentioned as well, such as visual complaints, headache, dizziness/imbalance as well as some less common indicators (Aroor, Singh, & Goldstein, 2017).

Regarding LDA, three topics for discussion were selected and the number of occurrences of a word in the entire corpus was measured (Hassabis, & Maguire, 2007). 'Arm' and 'speech' are the topics that the stroke survivors referred to the most, while the rest are vague as they contain words related to the general condition of the patients (e.g. 'hospital', 'stroke' and so on). In this way, the results of the tf we extracted are confirmed, since the symptoms selected by the LDA method are the most common topics discussed, too.

Based on our results, the speech and arm symptoms were common in the written transcriptions of the stroke survivors. The question of whether there is a relation to treatment cannot be answered by this study alone. As previous studies have shown, approximately three quarters of stroke patients do not interpret their symptoms correctly as representing a stroke and thus they do not seek for medical help after the symptoms' onset (Elkind, 2009; Kothari et al., 1997; Williams et al., 1997). Others have focused on the high levels of detection and diagnostic accuracy of stroke based on the FAST mnemonic (Harbison et al., 2003; Nor et al., 2004; Robinson et al. 2013). The results of our work provide preliminary evidence for the relevance of incorporating stroke symptoms in the evaluation of medical history intake. In the face of serious consequences of stroke, the examiner may find it difficult to transcribe and understand unintelligible speech. Thus, we created a computer-aided natural language processing methodology to help recognize relevant information in hopes of guiding the examiner to identify the importance of these symptoms and this is the first step in guiding rehabilitation: Memories of FAST symptoms should be recognized as part of the sequala of the stroke event and may serve as secondary prevention and self-care (Hassabis, & Maguire, 2007; Schacter, Addis, & Buckner, 2008; Wu, Cheng, & Chiou, 2017).

In conclusion, understanding the stroke symptoms in a transcribed dialogue, using computer-aided natural language processing techniques can be a starting point for additional research. In order to effectively educate stroke survivors to take care of themselves and improve rehabilitation systems globally, we first need to systematically continue to quantify what is reported about symptoms in the history intake which is the backbone of diagnosis. Future studies should examine the types of communication impairments (i.e. aphasia), as they relate to stroke symptoms, using standard natural language processing techniques in a larger sample of stroke survivors.

Ethical approval

This study was performed in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments. Subjects were from the rehab hospital Anagenissis and volunteers, approval number # 1102-2013. They signed a consent form and were debriefed at the end of the study. No reimbursement was given for participating in the study.

Acknowledgments

We would like to thank Chrysanthi Spyrou for helping with the analysis of the data.

Disclosure of potential conflicts of interest

On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

Abbe, A., Grouin, C., Zweigenbaum, P., & Falissard, B. (2016). Text mining applications in psychiatry: a systematic literature review. *International Journal of Methods in Psychiatric Research, 25*(2), 86-100. doi: 10.1002/mpr.1481

Aroor, S., Singh, R., & Goldstein, L. B. (2017). BE-FAST (Balance, Eyes, Face, Arm, Speech, Time): Reducing the Proportion of Strokes Missed Using the FAST Mnemonic. *Stroke*, *48*(2), 479-81. doi: 10.1161/STROKEAHA.116.015169

Bang, O. Y. (2017). Advances in biomarker for stroke patients: from marker to regulator. *Precision and Future Medicine*, *1*(1), 32-42. doi: 10.23838/pfm.2017.00052

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3*, 993-1022.

Damico, J. S., Müller, N., & Ball, M. J. (2010). *The handbook of language and speech disorders*. West Sussex: Wiley-Blackwell.

de Zubicaray, G., McMahon, K., Eastburn, M., & Pringle, A. (2006). Top down influences on lexical selection during spoken word production: A 4T fMRI investigation of refractory effects in picture naming. *Human brain mapping, 27*(11), 864-73. doi: 10.1002/hbm.20227

Dixon, G., Thornton, E. W., & Young, C. A. (2007). Perceptions of self-efficacy and rehabilitation among neurologically disabled adults. *Clinical Rehabilitation, 21*(3), 230–40.

Dombrowski, S. U., White, M., Machintosh, J. E., Gellert, P., Araujo-Soares, V., Thomson, R. G., et al (2015). The stroke 'Act FAST' campaign: Remembered but not understood?. *International Journal of Stroke, 10*(3), 324-330.

Elkind, M. S. V. (2009). Outcomes After Stroke: Risk of Recurrent Ischemic Stroke and Other Events. *The American Journal of Medicine, 122*, S7–S13.

Feigin, V. L., Forouzanfar, M. H., Krishnamurthi, R., Mensah, G. A., Connor, M., Bennett, D. A., et al. (2014). Global and regional burden of stroke during 1990-2010: findings from the Global Burden of Disease Study 2010. *Lancet (London, England)*, *383*(9913), 245–254. doi: 10.1016/s0140-6736(13)61953-4

Feigin, V. L, Norrving, B., & Mensah, G. A. (2017). Global Burden of Stroke. *Circulation Research, 120*(3), 439-448. doi: 10.1161/CIRCRESAHA.116.308413

Grawemeyer, B., Cox, R., & Lum, C. (2000). AUDIX: A knowledge-based system for speech-therapeutic auditory discrimination exercises. *Medical infobahn for Europe, 77,* 568-72. doi: 10.3233/978-1-60750-921-9-568

Harbison, J., Hossain, O., Jenkinson, D., Davis, J., Louw, S. J., Gary, A. (2003). Diagnostic accuracy of stroke referrals from primary care, emergency room physicians, and ambulance staff using the face arm speech test. *Stroke, 34*(1), 71-6.

Hassabis, D., & Maguire, E. A. (2007). Deconstructing episodic memory with construction. *Trends in Cognitive Sciences, 11*(7), 299-306. doi: 10.1016/j.tics.2007.05.001

Huang, S., Lv, Z., Guo, Y., Li, L. Zhang, Y. Zhou, L., et al. (2016). Identification of Blood Let-7e-5p as a Biomarker for Ischemic Stroke. *PLoS ONE*, *11*(10), e0163951. doi: 10.1371/journal.pone.0163951

Katan, M., & Elkind, M. S. V. (2018). The Potential Role of Blood Biomarkers in Patients with Ischemic Stroke: An Expert Opinion. *Clinical and Translational Neuroscience,* 1-7. doi: 10.1177/2514183X18768050

Kim, S. J, Moon, G. J., & Bang, O. Y. (2013). Biomarkers for stroke. *Journal of Stroke*, *15*(1), 27–37. doi: 10.5853/jos.2013.15.1.27

Kleindorfer, D. O., Miller, R., Moomaw, C. J., Alwell, K., Broderick, J. P., Khoury, J., et al. (2007). Designing a Message for Public Education Regarding Stroke: Does FAST Capture Enough Stroke?. *Stroke, 38*, 2864–8. doi: 10.1161/STROKEAHA.107.484329

Kothari, R., Sauerbeck, L., Jauch, E., Broderick, J., Brott, T., Khoury, J., & Liu, T. (1997). Patients' awareness of stroke signs, symptoms, and risk factors. *Stroke, 28*(10), 1871-5.

Kurowski, K., & Blumstein, S. E. (2016). Phonetic basis of phonemic paraphasias in aphasia: Evidence for cascading activation. *Cortex*, *75*, 193-203. doi: 10.1016/j.cortex.2015.12.005

Mangset, M., Tor, E. D., Forde, R., & Wyller, T. B. (2008). 'We're just sick people, nothing else': Factors contributing to elderly stroke patients' satisfaction with rehabilitation. *Clinical Rehabilitation, 22*(9): 825–35.

Moffatt, K., Pourshahid, G., & Baecker, R. M. (2017). Augmentative and alternative communication devices for aphasia: the emerging role of "smart" mobile devices. *Universal Access in the Information Society*, *16*, 115. doi: 10.1007/s10209-015-0428-x

Nor, A. M., McAllister, C., Louw, S. J., Dyker, A. G., Davis, M., Jenkinson, D., & Ford, G.A. (2004). Agreement between ambulance paramedic- and physician-recorded neurological signs with Face Arm Speech Test (FAST) in acute stroke patients. *Stroke, 35*(6), 1355-9.

Palmer, R., Hughes, H., & Chater, T. (2017). What do people with aphasia want to be able to say? A content analysis of words identified as personally relevant by people with aphasia. *PLoS One, 12*(3), e0174065. doi: 10.1371/journal.pone.0174065

Proios, H., Malatra, I., Christoudi, C., Willmes, K., Weniger, D., & Milonas, I. (2006). Development and validation of the Aachen Aphasia Test in the Hellenic Language, paper delivered at the International Neuropsychological Society (INS/SVNP/GNP) Meeting "From Plasticity to Rehabilitation", Zurich, Switzerland.

Ranney, M. L., Choo, E. K., Cunningham, R. M., Spirito, A., Thorsen, M., Mello, M. J., & Morrow, K. (2014). Acceptability, language, and structure of text message-based behavioral interventions for high-risk adolescent females: a qualitative study. *Journal of Adolescent Health, 55*(1), 33–40. doi: 10.1016/j.jadohealth.2013. 12.017

Robinson, T. G., Reid, A., Haunton, V. J., Wilson, A. & Naylor, A.

R. (2013). The face arm speech test: does it encourage rapid recognition of important stroke warning symptoms? *Emergency Medicine Journal, 30*(6), 467-71. doi: 10.1136/emermed-2012-201471

Roundtree, A. K. (2018). From Engineers' Tweets: Text Mining Social Media for Perspectives on Engineering Communication. In: *IEEE International Professional Communication Conference (ProComm),* Toronto, Canada, 22-25 July, pp. 6-15. IEEE.

Salton, G., & Buckley, C. (1998). Term-weighting approaches in automatic text retrieval. *Information processing & management,* 24(5), 513-23. doi: 10.1016/0306-4573(88)90021-0

Schacter, D. L., Addis, D. R., & Buckner, R. L. (2008). Episodic Simulation of Future Events. *Annals of the New York Academy of Sciences, 1124*, 39-60. doi: 10.1196/annals.1440.001

Sponsler, J. L., & Burkhart, J. C. (2016). Aphasia Dx: An Expert Diagnostic System for Aphasia. In: *Telehealth and Assistive Technology / 847: Intelligent Systems and Robotics*, Zurich, Switzerland, January. doi: 10.2316/P.2016.847-005

Vannobel, J., & Toulotte, J. M. (1992). An Expert System For Aphasia Treatment Using Exercises Concerning Alphabetically Written Numeral Forms. In: *14th annual International Conference of the IEEE Engineering in Medicine and Biology Society,* Paris, France, paper no. 3, pp. 922-3. doi.org/10.1109/IEMBS.1992.5761309

Ward, A., Gbadebo, A., & Baruah, B. (2015). Using job advertisements to inform curricula design for the key global technical challenges. In: *International Conference on Information Technology Based Higher Education and Training (ITHET),* Caparica, Lisbon, Portugal, 11-13 June, pp. 1-6. IEEE.

Williams, L. S., Bruno, A., Rouch, D., & Marriott, D. J. (1997). Stroke patients' knowledge of stroke. Influence on time to presentation. *Stroke*; *28*(5), 912-5.

Worrall, L., Sherratt, S., Rogers, P., Howe, T., Hersh, D. Ferguson, A., & Davidson, B. (2011). What people with aphasia want: Their goals according to the ICF. *Aphasiology*, *3*, 309-22. doi: 10.1080/02687038.2010.508530

Wu, W. H., Cheng, W. & Chiou, W. B. (2017). Episodic future thinking about the ideal self induces lower discounting, leading to a decreased tendency toward cheating. *Frontiers in Psychology, 8,* Article ID 287. doi: 10.3389/fpsyg.2017.00287

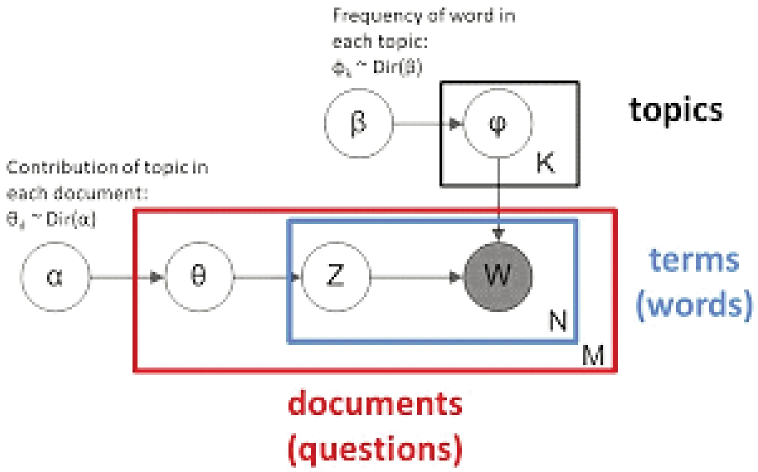

Fig1 LDA topic selection procedure

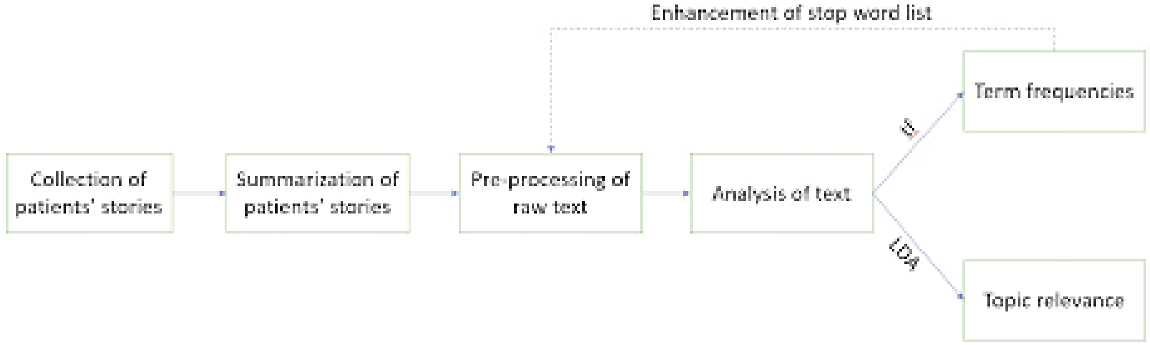

Fig1 Text analysis procedure

this method as well.

Word	Freq.	Word	Freq.	Word	Freq.	Word	Freq.
speech	43	incident	8	pain	5	dizziness	3
weakness	24	stroke	7	after	5	ability	2
felt	19	confusion	7	sense	5	headache	2
arm	18	transferred	6	symptom	4	surgery	2
hospital	11	paralysis	6	lost	4	claims	2
leg	11	started	5	followed	4	instability	2
remember	9	numbness	5	sickness	4	alone	2
intense	9	reports	5	problem	4	immediately	2
suddenly	8	eye	5	body	3	headache	2

Table 1 Single word frequency

Word	Freq.	Word	Freq.	Word	Freq.
inability talk	52 52	Weaken end	8	that fell	4
weakness arm	26	Walk well	8	mentions is	4
can not	24	Incident started	4	eye weaken	4
transferred hospital	18	arm started	4	Spastic movement	4
speech confusion	12	first symptom	4	suddenly felt	2
started weaken	10	had control	4	Lose consciousness	2
didn't have	8	started feeling	4	senses arm	2
clamp lightly	8	leg started	4	profound weakness	2
arm leg	8	according saying	4	Could still	2

***Table 2** Double word frequency*

Word	Freq.	Word	Freq.
can not talk	28	psychology happened stroke	6
lose senses arm	24	strong headache dizziness	6
started weaken finally	14	has happened alteration	4
felt eye weaken	12	had surgery heart	4
always problem diabetes	12	had bleeding stomach	4
still clamp lightly	8	consciousness transferred hospital	6
make spastic movement	6	always had problem	2
doesn't remember anything	6	Clamp paralysis leg	6
speech confusion thinking	6	Finally jaw stomped	2

***Table 3** Three-word frequency*

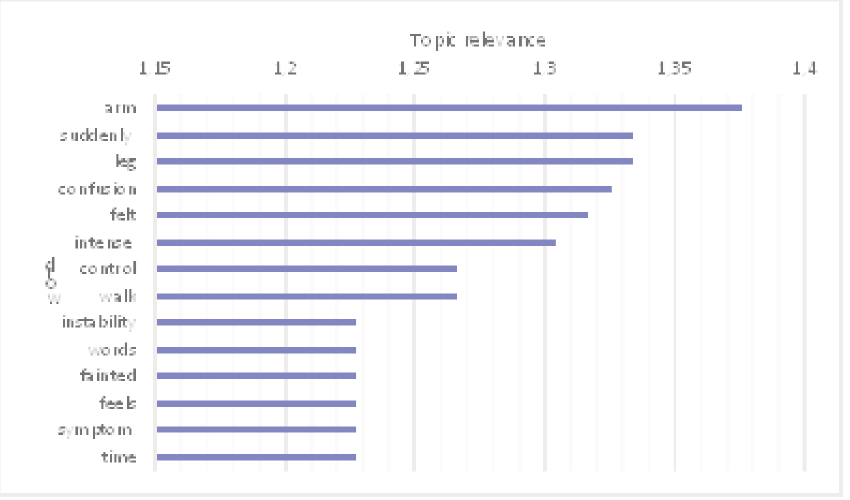

Fig1 Topic 1 [arm]

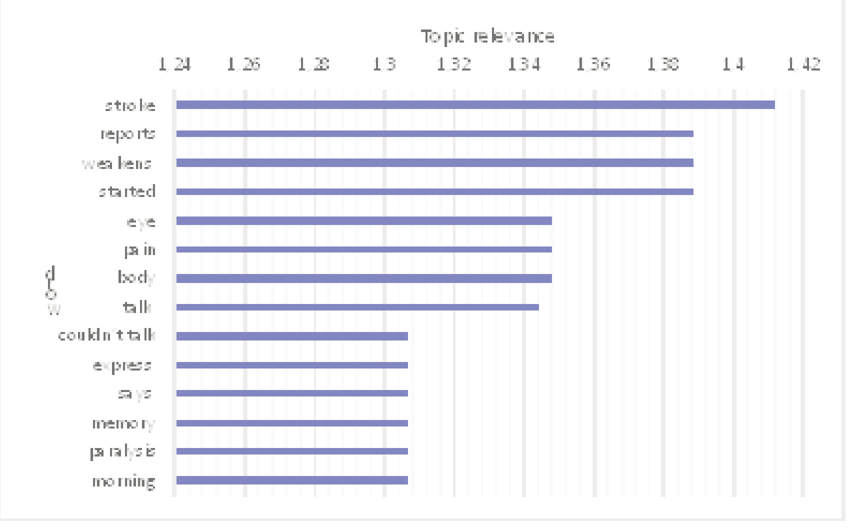

Fig2 Topic 2 [speech]

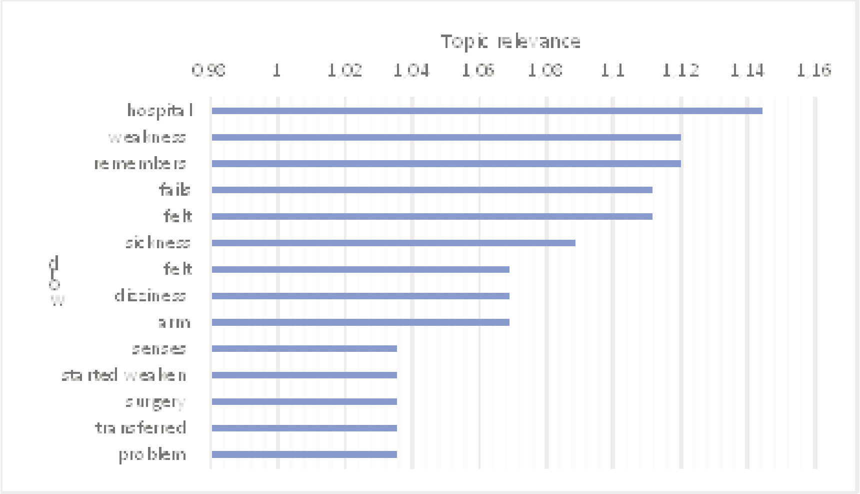

***Fig1** Topic 3 [stroke issues in general]*