Detecting Stress Reasons Using Automatic Topic Modeling

[1]Isha Mahajan, [2]Vinayak Jena, [3]Sanjaya Gurram, [4]Prof. Mohini Chaudhari

[1,2,3]Research Scholar, [4]Assistant Professor, Vidyalankar Institute of Technology, Mumbai, India.

[1]contactishamahajan@gmail.com, [2]vinayakjena1997@gmail.com, [3]gurramsanjaya@gmail.com, [4]mohinichaudhari89@gmail.com

Abstract — Stress and depression have been dubbed as critical issues contributing to weakening of physical and mental health. Cigna TTK Health Insurance conducted a survey according to which about 89% of the population in India suffers from stress compared to the global average of 86%. World Health Organization has pinpointed stress as "health epidemic of 21st century". Mental health is imperative to maintain a sound life as it influences the daily routine and not only personal but social behaviour as well. Therefore, it is necessary to detect stress as health is essential for overall growth, productivity and development of the society. We propose a system to automatically detect reasons of stress in people's daily life by applying Automatic Topic Modeling using Latent Dirichlet Allocation (LDA) on data provided by J. Schler, M. Koppel, S. Aragmon and J. Pennebaker's Blog Authorship Corpus which we will refine to suit our requirements.

Keywords — Automatic Topic Modeling, Detecting Stress Reasons, LDA, Natural Language Processing.

I. INTRODUCTION

Psychological stress refers to emotional and physiological reactions experienced when someone confronts a situation in which the demands go beyond their coping mechanisms. Hans Selye, credited as being father of stress gave the first ever definition of psychological stress in the year 1936 which states stress as "the non-specific response of the body to any demand for changes". The circumstances or situations that cause stress are known as stressors. The cause of stress depends on our perception of it. Something that puts pressure on us may not be that worrisome to others. Stressors are usually thought of by people as being negative, like an overload of work or excessive worrying, although anything that puts pressure on people can be worrisome and stressful; this includes positive events such as getting a promotion or getting married. Frequent external causes of stress include but are not limited to workplace problems, overload of responsibilities, financial problems, etc and frequent internal causes of stress include but are not limited to negative self-talk, excessive worrying, lot of expectations [1].

Stress can directly increase heart rate, blood flow, worsen asthma and diabetes, cause lack of sleep, increase susceptibility to infections and cardiovascular diseases. Stress exists in two forms: Acute stress and Chronic stress. Acute stress is short-lived stress that exists for the time being for which human body is designed to recover from.

It's an instant reaction to a new challenge or demand that is activated instantly like a fight or flight response while chronic stress on the other hand is when acute stress can't be resolved and begins to increase continuously and persists for a long term. Rich and Bonner found in a stress-vulnerability model that negative life events and stress accounted for 30% of the variance in suicidal ideation. Chronic stress is connected with higher rate of depression and anxiety.

Chronic stress results in reduced serotonin, increase of hormones like cortisol which is the stress hormone as well as other neurotransmitters in brain including dopamine, which is linked to depression. These chemical systems allow expression of normal moods and emotions and manage biological processes like digestion, sleep, energy and so forth however, once the stress response fails to stop and reset after a troublesome situation has passed, it can lead to depression in vulnerable people [2]. Therefore detecting stress at an early stage before it becomes a severe problem is extremely necessary.

In this study we focus on detecting the reasons of stress in people's daily life by applying Automatic Topic Modeling using Latent Dirichlet Allocation (LDA) on data provided by J. Schler, M. Koppel, S. Aragmon and J. Pennebaker's Blog Authorship Corpus. The Blog Authorship Corpus has a total of 19,320 bloggers with a total of 681,288 posts gathered from blogger.com which will be further refined to suit our requirements [3].

Every document consists of multiple different topics. Topics are groups of words that occur frequently in the document. Topic modeling algorithms are statistical methods which analyze word-topic distribution and topic-document distribution to discover the abstract themes or say topics that run through them, how the themes are interconnected and change over time [4].

DOI : 10.18231/2454-9150.2018.1409

II. Related Work

There have been many techniques developed to detect stress with the help of data collected using physiological sensors or face-to-face interviews conducted by psychologists which usually rely on the active individual participation hence, it becomes significant to detect stress timely for proactive care [5].

Social media platform has become an attraction for people to express their feelings and daily life events. Several research papers have attempted to identify stress levels, depression and other mental health disorders through natural language processing of social media however not much work has been done comparatively towards discovering reasons behind the stress [6].

Wang Rui et al. (2014) proposed an android app ‘StudentLife’ to utilize automatic and consistent Smartphone sensing to evaluate emotional well-being (example - stress, loneliness, depression), academic standing and behavioural trends (example- how sleep, stress, eating habits, and so forth change because of tasks at hand or college workload like midterms, assignments, finals as the term advances) of a student body [7].

Philip Resnik et al. (2013) studied the benefits of topic modeling in text analysis for neuroticism and depression as a strongly associated personality measure. To provide baseline features - Pennebaker’s LIWC lexicon was used. They show that uncomplicated and straightforward topic modeling using Latent Dirichlet Allocation yields explicable, psychologically relevant topics or themes that add value in prediction of clinical assessments [6].

De Choudhury et al. (2013) used crowdsourcing to build an extensive dataset of Twitter posts along with questionnaires for psychological evaluation and a survey which had each user’s history of clinical depression. To detect depression, supervised learning models were built with a basic set of features [8].

B. Padmaja et al. (2018) proposed a system to detect cognitive stress levels using sensor technology in the form of a physical activity tracker device developed by FITBIT where each stressor’s effect was evaluated using logistic regression and then a combined model was built and assessed using variants of ordinal logistic regression models using logit, probit and complementary log-log [9].

Yuan Gong et al. (2017) proposed a topic modeling approach demonstrating a way to perform context-aware analysis for the 2017 Audio/Visual Emotion Challenge that requested participants to make a model that could predict depression levels based on the text files, audio and video of an interview [10].

A lot of work has been done to detect stress and depression and their levels but our system would detect the reasons behind the stress that people are facing using analysis of textual data instead of audio or visual. LDA is being used as the topic modeling algorithm as it is better at identifying coherent topics due to its consistency. As the study of Philip Resnik et al. who demonstrated that LDA yields psychologically relevant themes, LDA is a better choice for this system.

III. Proposed System

Mental health is important to maintain a healthy life as it affects the daily routine and not only personal but social behaviour as well. Therefore, it is necessary to detect stress as health is essential for overall growth, productivity and development of the society. Current systems use data from Smartphone, activity trackers or use Topic modeling to detect if a person is stressed or not and to detect the level of stress.

A topic consists of combination of words occurring frequently. Topic model can connect words with similar meanings using contextual clues and distinguish between multiple meanings and uses of words. We propose a system to automatically detect reasons of stress in people’s daily life by applying Automatic Topic Modeling using Latent Dirichlet Allocation (LDA) to train the data provided by J. Schler, M. Koppel, S. Aragmon and J. Pennebaker’s Blog Authorship Corpus which we will refine to suit our requirements.

Initially cleaning and preprocessing of the data would be done by performing Tokenization, removal of stop words and stemming. The next step would be running the LDA model where number of topics would be fixed. Then each topic would be assigned an associated probability distribution over words using a random variable. Another random variable would assign each document a probability distribution over topics. These are the document-topic and Topic-word matrices which would further need improvement as the values had been assigned randomly using random variables.

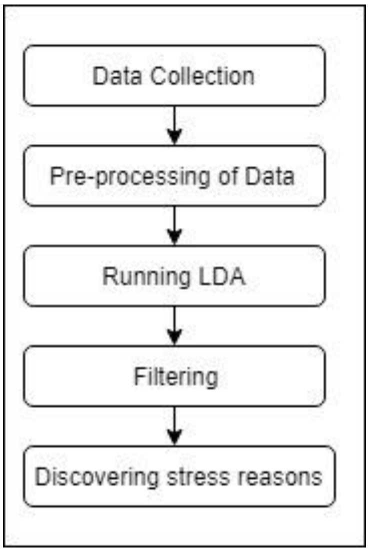


Fig. 1. Block diagram of stress reasons detection system

Gibbs Sampling technique will be used to improve the matrices. It goes through every word of every document and updates the current topic-word assignment with a new assignment. After huge number of iterations a stable state is achieved where all the assignments are adequate and leads to a convergence point. This would give us a number of abstract topics or themes in the documents. Out of all these topics we would select the topic which is most related to reasons of stress, this topic would contain a set of words i.e. reasons of stress which is the expected result.

IV. Methodology

The concept of Topic Modeling is that a document is composed of a variety of different topics. Topic Modeling is an unsupervised approach that helps discover these latent or abstract topics present in the document. Topics are groups of words that occur frequently in the document. Many techniques have been developed to obtain topic models. The one proposed in our system is Latent Dirichlet Allocation.

LDA assumes that documents are created from a mixed number of themes or topics and then the topics generate words primarily dependent on their probability distribution. Given a corpus, LDA backtracks and endeavours to make sense of which topics would produce the set of documents in the first place [4].

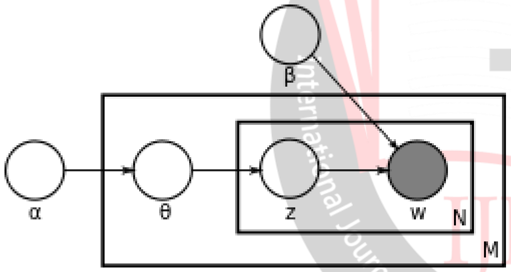


Fig. 2. LDA Model [11]

Fig. 2 shows the LDA model where parameters of LDA are Alpha and Beta Hyperparameters, number of topics and number of topic terms. α is the per document topic distribution, β is the per topic word distribution, for document *m* - θm is the topic distribution, φm is the word distribution for topic *k*, Zmn is the topic for the *n*th word in document *m*, and Wmn is the definite word.

A. Data Collection

The database to be used in this system is provided by J. Schler, M. Koppel, S. Aragmon and J. Pennebaker's Blog Authorship Corpus which we will refine to suit personal writing blog contents and remove all those blogs that deal with information on technological advances and other topics that are unrelated to this system. The Blog Authorship Corpus has a total of 19,320 bloggers with a total of 681,288 posts gathered from blogger.com.

B. Pre-processing of Data

The blogs are combined together to form a corpus. Cleaning or pre-processing is a vital step before any text mining task. Tokenization, removal of stop words and stemming is to be done. Here, tokenization is breaking down sequence of strings into words. Removal of stop words avoids generating huge number of topics filled with words like 'the', 'and', 'of', 'to', etc. Sometimes a corpus results in many topics containing some common stop words which can be avoided by adding corpus specific stop words. All these steps together constitute the Pre-processing of the corpus.

C. Latent Dirichlet Allocation

LDA assumes that topics are generated first and then the documents are generated after that. As documents are made of various different topics so a fixed number of k topics are chosen. The corpus containing blog posts is represented in the form of a document-term matrix where the matrix demonstrates a corpus of N documents D1,D2,...,Dn and size of M words W1, W2,...,Wn.

The value of cell i,j provides the frequency count of word Wj occurring in document Di. LDA bifurcates this document-term matrix to produce two lower dimensional matrices – M1 and M2 as shown in Fig. 3 where M1 is a document-topic matrix with dimension (N,K) and M2 is a Topic-word matrix with dimension (K,M) where M is the vocabulary size, K is the fixed number of topics and N is the number of documents [12].

As each and every word in the document is randomly assigned to any one of the topics, both the matrices give us topic-word and document-topic distributions. However, improving these distributions is the primary goal of LDA. To improve these matrices, LDA uses sampling techniques so we'll use collapsed Gibbs sampling to learn the themes or topics and the topic representations of every document.

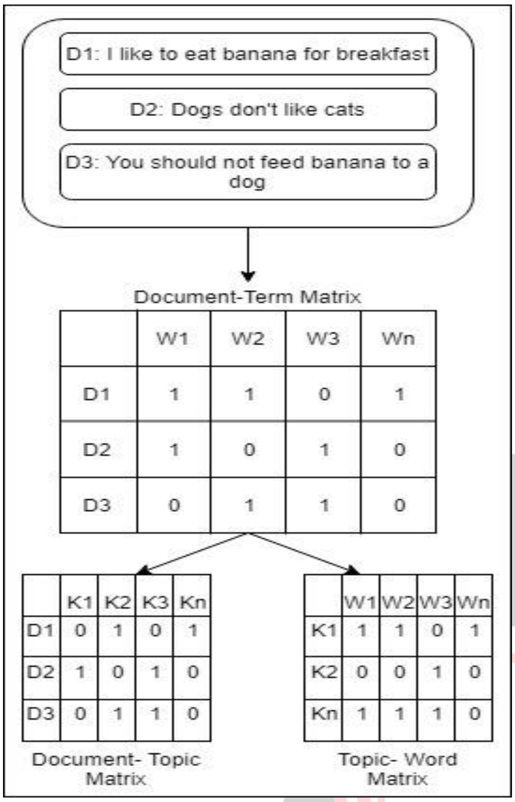
Therefore to improve the matrices, we'll go through each word in a document D and for each topic t, two things will be computed: 1) P(topic t | document D) i.e the total of all the words in document D which are presently assigned to topic t. 2) P(word w | topic t) i.e the proportion of assignments to topic t that originate from word w.

Based on the product of the above two probabilities i.e P(topic t | document D * P(word w | topic t) , a new topic t is chosen and w is reassigned this new topic (this is basically the likelihood that word w was produced by topic t, so it bodes well to update the current word's topic with this probability). In simpler words, here we're taking into account the assumption that all topic assignments apart from the current word are precise, and therefore we would be resampling the assignment of the current word utilizing our model of how the documents are produced.

After performing the last step repeatedly a huge number of times, we'll ultimately reach a stable state where our assignments are adequate. This would be the convergence point. Then these assignments will be used to evaluate the

DOI : 10.18231/2454-9150.2018.1409

topic assortments of every document (by tallying the proportion of words assigned to each topic inside that particular document) and the words affiliated to each topic (by tallying proportion of the words assigned to each topic overall).



D1: I like to eat banana for breakfast

D2: Dogs don't like cats

D3: You should not feed banana to a dog

Document-Term Matrix

	W1	W2	W3	Wn
D1	1	1	0	1
D2	1	0	1	0
D3	0	1	1	0

	K1	K2	K3	Kn
D1	0	1	0	1
D2	1	0	1	0
D3	0	1	1	0

Document- Topic Matrix

	W1	W2	W3	Wn
K1	1	1	0	1
K2	0	0	1	0
Kn	1	1	1	0

Topic- Word Matrix

Fig. 3. Example of bifurcation of Document-Term Matrix

D. Filtering

LDA will generate different number of topics. All these topics are latent or abstract topics in the refined corpus. Out of the entire list of topics, one topic will be the most relative topic regarding stress reasons. This topic would contain a set of words that show the reasons of stress people have faced based on what they had written in the blogs.

V. Conclusion

In this paper the idea of detecting stress reason using LDA for Automatic Topic Modeling is presented. In today’s era social media has become an attraction for people to talk about their feelings and day to day life. Stress percentage is on the increase, due to the openness of people on social media, we have been given a window of opportunity to detect reasons of stress. LDA provides flexibility and is good at identifying coherent topics as it is consistent. Fixing the number of topics ought to be dealt with properly, if there are less or considerably more number of topics then the result probably won’t be that accurate.

References

[1] "Stress Symptoms, Signs, and Causes", *HelpGuide.org*, 2019.[Online].Available:https://www.helpguide.org/articles/stress/stress-symptoms-signs-and-causes.htm.

[2] Karen Bruno, "Stress and Depression", *WebMD*, 2018. [Online].Available:https://www.webmd.com/depression/features/stress-depression#1.

[3] J. Schler, M. Koppel, S. Argamon and J. Pennebaker (2006). Effects of Age and Gender on Blogging in *Proceedings of 2006 AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs.*

[4] D. Blei, "Probabilistic topic models", *Communications of the ACM*, vol. 55, no. 4, p. 77, 2012.

[5] Hajera, S. and Ali, M.M., “A Comparative Analysis of Psychological Stress Detection Methods”, *ijcem*, vol. 21, 2018.

[6] Resnik, Philip, Anderson Garron, and Rebecca Resnik. "Using topic modeling to improve prediction of neuroticism and depression in college students." *Proceedings of the 2013 conference on empirical methods in natural language processing*. 2013.

[7] Wang, Rui, et al. "StudentLife: assessing mental health, academic performance and behavioral trends of college students using smartphones." *Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing*. ACM, 2014.

[8] De Choudhury, Munmun, et al. "Predicting depression via social media." *ICWSM* 13 (2013): 1-10.

[9] B. Padmaja, V. Prasad and K. Sunitha, "Machine Learning Approach for Stress Detection using Wireless Physical Activity Tracker", *International Journal of Machine Learning and Computing*, vol. 8, no. 1, pp. 33-38, 2018.

[10] Gong, Yuan, and Christian Poellabauer. "Topic Modeling Based Multi-modal Depression Detection." *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*. ACM, 2017.

[11] "Latent Dirichlet allocation", *En.wikipedia.org*, 2019. [Online].Available:https://en.wikipedia.org/wiki/Latent_Dirichlet_allocation. [Accessed: 30- Jan- 2019].

[12] D. Nghia, "Beginners Guide to Topic Modeling in Python", *Nghia's Blog*, 2019. [Online]. Available: https://duongtrungnghia.wordpress.com/2017/02/10/beginners-guide-to-topic-modeling-in-python/.