12-2018

Data mining approach to the detection of suicide in social media: A case study of Singapore

Jane H. K. SEAH
Singapore Management University, jane.seah.2016@smu.edu.sg

Kyong Jin SHIM
Singapore Management University, kjshim@smu.edu.sg

Citation

SEAH, Jane H. K. and SHIM, Kyong Jin. Data mining approach to the detection of suicide in social media: A case study of Singapore. (2018). *2018 IEEE International Conference on Big Data (Big Data): Seattle, WA, December 10-13: Proceedings*. 5442-5444.

Available at: https://ink.library.smu.edu.sg/sis_research/4340

2018 IEEE International Conference on Big Data (Big Data)
pp. 5442-5444. DOI: 10.1109/BigData.2018.8622528

Data Mining Approach to the Detection of Suicide in Social Media: A Case Study of Singapore

Jane H. K. Seah
School of Information Systems
Singapore Management University
Singapore
jane.seah.2016@sis.smu.edu.sg

Kyong Jin Shim
School of Information Systems
Singapore Management University
Singapore
kjshim@smu.edu.sg

***Abstract*—In this research, we focus on the social phenomenon of suicide. Specifically, we perform social sensing on digital traces obtained from Reddit. We analyze the posts and comments in that are related to depression and suicide. We perform natural language processing to better understand different aspects of human life that relate to suicide.**

Keywords—suicide, suicide detection, social media, data mining

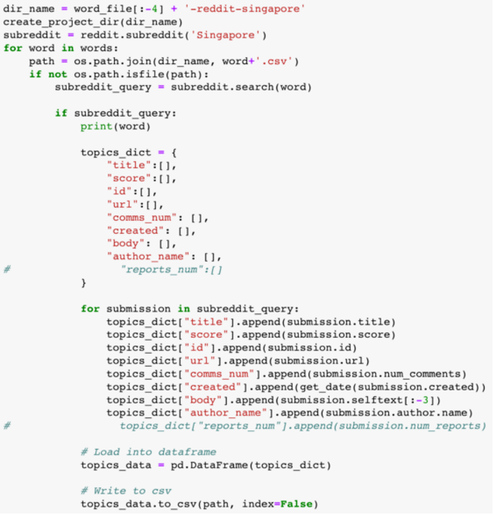
I. Introduction

Suicide is a global phenomenon in all regions of the world. The World Health Organization reports [1] that some 800,000 lives are lost annually due to suicide. A recent article [5] reports that the global suicide rate in developed countries rose since 1999 and by the year 2010, suicide became the major reason behind unnatural death. Singapore is not an exception. As population ageing trend continues, more cases of elderly suicide are being reported in Singapore [2]. Also, the number of suicide-related deaths is reported to be far exceeding that of transport-related accidents [3]. The same report shows that especially for the younger age group of 10 years old to 29 years old, suicide is known to be the leading cause of their death. Just last year, a suicide case involving a young student [6] was reported in Singapore. According to the article [6], extreme stress caused by exams is suspected to have caused the suicide.

With over 900 active social media platforms globally as of today, platforms such as Facebook, Twitter and Reddit have seen a drastic increase in active conversations about various topics and issues globally. Such digital traces provide valuable insights into human emotional states and feelings. While offline methods such as hotlines and in-person counselling exist to capture human emotional states and feelings, social media platforms provide rich digital traces that can be analyzed to detect suicidal intentions using data mining techniques.

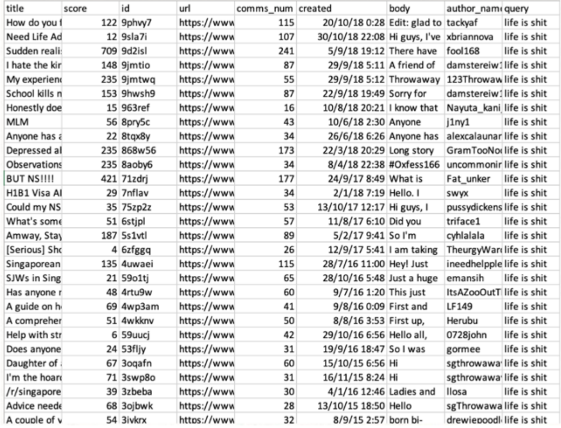
II. Dataset

In this study, we investigate depression and suicide-related conversations in Singapore. We performed data crawling from Reddit using their API. After obtaining an API key, we wrote Python scripts to search and crawl posts and comments from Subreddit "/r/Singapore." We use PRAW [7], a popular Python package and a Reddit API wrapper. Using the pre-defined and customized list of key terms related to depression and suicide, the Python scripts then used these terms to query Reddit (Figure 1). The crawled data are stored as CSV files (Figure 2).



```
dir_name = word_file[:-4] + '-reddit-singapore'
create_project_dir(dir_name)
subreddit = reddit.subreddit('Singapore')
for word in words:
    path = os.path.join(dir_name, word+'.csv')
    if not os.path.isfile(path):
        subreddit_query = subreddit.search(word)

        if subreddit_query:
            print(word)

            topics_dict = {
                "title":[],
                "score":[],
                "id":[],
                "url":[],
                "comms_num": [],
                "created": [],
                "body": [],
                "author_name": [],
#                 "reports_num":[]
            }

            for submission in subreddit_query:
                topics_dict["title"].append(submission.title)
                topics_dict["score"].append(submission.score)
                topics_dict["id"].append(submission.id)
                topics_dict["url"].append(submission.url)
                topics_dict["comms_num"].append(submission.num_comments)
                topics_dict["created"].append(get_date(submission.created))
                topics_dict["body"].append(submission.selftext[:-3])
                topics_dict["author_name"].append(submission.author.name)
#                 topics_dict["reports_num"].append(submission.num_reports)

            # Load into dataframe
            topics_data = pd.DataFrame(topics_dict)

            # Write to csv
            topics_data.to_csv(path, index=False)
```

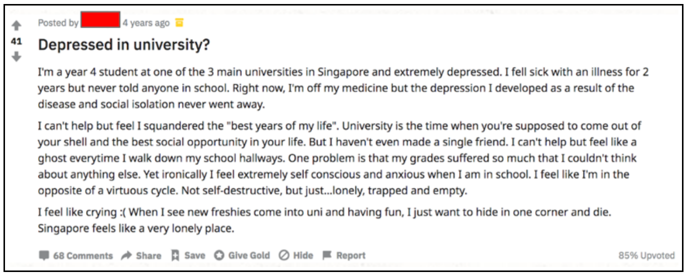
Fig. 1. Reddit Crawling Script Written in Python



title	score	id	url	comms_num	created	body	author_nam	query
How do you f	122	9phvy7	https://www	115	20/10/18 0:28	Edit: glad to	tackyaf	life is shit
Need Life Ad	12	9sla7i	https://www	107	30/10/18 22:08	Hi guys, I've	xbriannova	life is shit
Sudden reali:	709	9d2isl	https://www	241	5/9/18 19:12	There have	fool168	life is shit
I hate the kir	148	9jmtio	https://www	87	29/9/18 5:11	A friend of	damstereiw1	life is shit
My experienc	235	9jmtwq	https://www	55	29/9/18 5:12	Throwaway	123Throwaw	life is shit
School kills n	153	9hwsh9	https://www	87	22/9/18 19:49	Sorry for	damstereiw1	life is shit
Honestly doe	15	963ref	https://www	16	10/8/18 20:21	I know that	Nayuta_kani	life is shit
MLM	56	8pry5c	https://www	43	10/6/18 2:30	Anyone	j1ny1	life is shit
Anyone has a	22	8tqx8y	https://www	34	26/6/18 6:26	Anyone has	alexcalaunar	life is shit
Depressed al	235	868w56	https://www	173	22/3/18 20:29	Long story	GramTooNo	life is shit
Observations	235	8aoby6	https://www	34	8/4/18 22:38	#Oxfess166	uncommonir	life is shit
BUT NS!!!!	421	71zdrj	https://www	177	24/9/17 8:49	What is	Fat_unker	life is shit
H1B1 Visa Al	29	7nflav	https://www	34	2/1/18 7:19	Hello. I	swyx	life is shit
Could my NS	35	75zp2z	https://www	53	13/10/17 12:17	Hi guys, I	pussydickens	life is shit
What's some	51	6stjpl	https://www	57	11/8/17 6:10	Did you	triface1	life is shit
Amway, Stay	187	5s1vtl	https://www	89	5/2/17 9:41	So I'm	cyhlalala	life is shit
[Serious] Sho	4	6zfggq	https://www	26	12/9/17 5:41	I am taking	TheurgyWar	life is shit
Singaporean	135	4uwaei	https://www	115	28/7/16 11:00	Hey! Just	ineedhelpple	life is shit
SJWs in Sing	21	59o1tj	https://www	65	28/10/16 5:48	Just a huge	emansih	life is shit
Has anyone r	48	4rtu9w	https://www	60	9/7/16 1:20	This just	ItsAZooOutT	life is shit
A guide on h	69	4wp3am	https://www	41	9/8/16 0:09	First and	LF149	life is shit
A comprehen	51	4wkknv	https://www	50	8/8/16 3:53	First up,	Herubu	life is shit
Help with str	6	59uucj	https://www	42	29/10/16 6:56	Hello all,	0728john	life is shit
Does anyone	24	53fljy	https://www	31	19/9/16 18:47	So I was	gormee	life is shit
Daughter of a	67	3oqafn	https://www	60	15/10/15 6:56	Hi	sgthrowawa	life is shit
I'm the hoard	71	3swp8o	https://www	31	16/11/15 8:24	Hi	sgthrowawa	life is shit
/r/singapore	39	3zbeba	https://www	30	4/1/16 12:46	Ladies and	llosa	life is shit
Advice neede	68	3ojbwk	https://www	28	13/10/15 18:50	Hello	sgThrowawa	life is shit
A couple of v	54	3ivkrx	https://www	32	8/9/15 2:57	born bi-	drewiepoodl	life is shit

Fig. 2. Depression and Suicide-Related Posts and Comments Crawled from Reddit

Figure 3 shows a depression-related post in Reddit that our script crawled. 385 posts and over 21,000 comments mentioning depression and suicide-related terms were retrieved from Reddit. The data span the Year 2010 through the Year 2018. We used the suicide-related terms from a previous study [4], which includes terms indicative of depression and suicide (e.g. "feel useless", "hang myself", known suicide methods, prescription drug names, etc.).



41

Posted by [redacted] 4 years ago

Depressed in university?

I'm a year 4 student at one of the 3 main universities in Singapore and extremely depressed. I fell sick with an illness for 2 years but never told anyone in school. Right now, I'm off my medicine but the depression I developed as a result of the disease and social isolation never went away.

I can't help but feel I squandered the "best years of my life". University is the time when you're supposed to come out of your shell and the best social opportunity in your life. But I haven't even made a single friend. I can't help but feel like a ghost everytime I walk down my school hallways. One problem is that my grades suffered so much that I couldn't think about anything else. Yet ironically I feel extremely self conscious and anxious when I am in school. I feel like I'm in the opposite of a virtuous cycle. Not self-destructive, but just...lonely, trapped and empty.

I feel like crying :(When I see new freshies come into uni and having fun, I just want to hide in one corner and die. Singapore feels like a very lonely place.

Fig. 3. A Depression-Related Reddit Post (the user's name is masked).

Prior to programmatically performing data cleaning and transformation, our human researchers manually inspected the crawled data to ensure relevancy of the data. Some 5% of the data contained Chinese characters (as Chinese is widely spoken in Singapore). We engaged human translators to translate Chinese phrases and words into English. Initially, we used freely available online translation services, however, the quality of the translation was deemed poor. A majority of the translated content turned out to be irrelevant to our topic (e.g. comments related to recent political elections), and hence, we removed it from our analysis.

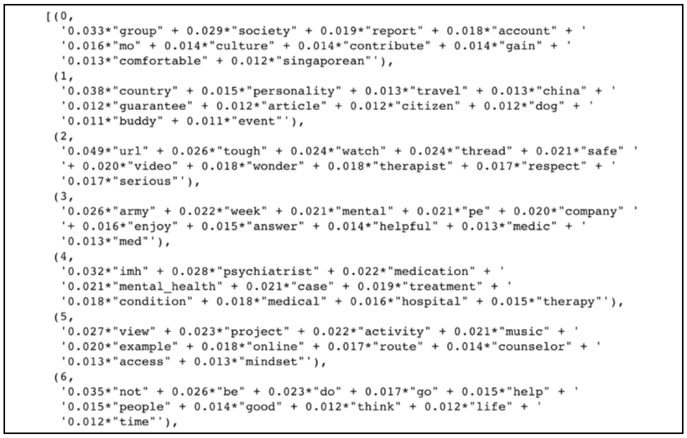
Our researchers also spent a substantial amount of time expanding our corpus with acronyms or terms previously unknown. For instance, "pes" or "physical employment status" was mentioned in some of the posts. This term is specific to Singapore's National Service. Physical Employment Status (PES) is a required physical fitness exam for National Service registrants, enlisted soldiers and reservists. As it appeared along with depression and suicide indicating terms, we took note of this and added it to our corpus.

We further enhanced this corpus with new terms such as new drug names associated with treatment of depression, other suicide methods, and mostly local terms such as "imh" (Institute of Mental Health in Singapore) and terms related to Singapore's National Service. We included these new terms in our customized corpus based on term frequency in the data that our study obtained.

Further data cleaning and transformation involved removal of certain stop words, Unicode characters, and hyperlinks. Upon manual data cleaning and inspection, next we performed stemming, lemmatization, spell checking prior to performing analyses.

III. FINDINGS & ANALYSIS

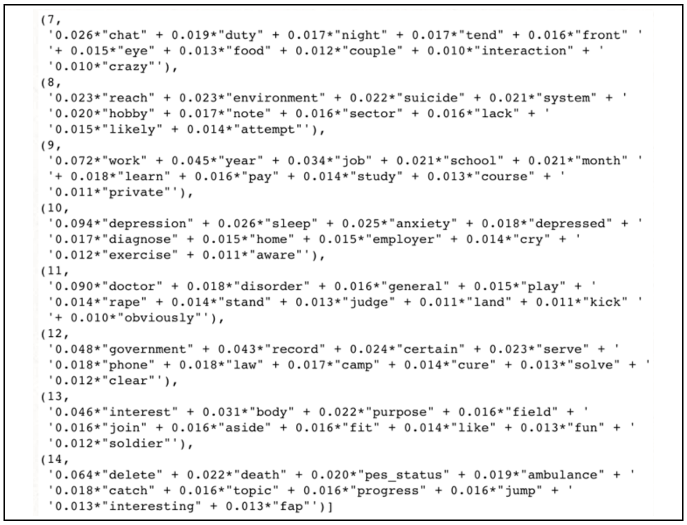
We performed topic modeling using Latent Dirichlet Allocation (LDA) [8] which automatically discovers topics. We used Python Gensim package [9] for topic modelling.



```
[(0,
  '0.033*"group" + 0.029*"society" + 0.019*"report" + 0.018*"account" + '
  '0.016*"mo" + 0.014*"culture" + 0.014*"contribute" + 0.014*"gain" + '
  '0.013*"comfortable" + 0.012*"singaporean"'),
 (1,
  '0.038*"country" + 0.015*"personality" + 0.013*"travel" + 0.013*"china" + '
  '0.012*"guarantee" + 0.012*"article" + 0.012*"citizen" + 0.012*"dog" + '
  '0.011*"buddy" + 0.011*"event"'),
 (2,
  '0.049*"url" + 0.026*"tough" + 0.024*"watch" + 0.024*"thread" + 0.021*"safe" '
  '+ 0.020*"video" + 0.018*"wonder" + 0.018*"therapist" + 0.017*"respect" + '
  '0.017*"serious"'),
 (3,
  '0.026*"army" + 0.022*"week" + 0.021*"mental" + 0.021*"pe" + 0.020*"company" '
  '+ 0.016*"enjoy" + 0.015*"answer" + 0.014*"helpful" + 0.013*"medic" + '
  '0.013*"med"'),
 (4,
  '0.032*"imh" + 0.028*"psychiatrist" + 0.022*"medication" + '
  '0.021*"mental_health" + 0.021*"case" + 0.019*"treatment" + '
  '0.018*"condition" + 0.018*"medical" + 0.016*"hospital" + 0.015*"therapy"'),
 (5,
  '0.027*"view" + 0.023*"project" + 0.022*"activity" + 0.021*"music" + '
  '0.020*"example" + 0.018*"online" + 0.017*"route" + 0.014*"counselor" + '
  '0.013*"access" + 0.013*"mindset"'),
 (6,
  '0.035*"not" + 0.026*"be" + 0.023*"do" + 0.017*"go" + 0.015*"help" + '
  '0.015*"people" + 0.014*"good" + 0.012*"think" + 0.012*"life" + '
  '0.012*"time"'),
```

Fig. 4. Some of the Topics Discovered by LDA

Figure 4 and Figure 5 show the topics the LDA algorithm discovered from our Reddit dataset.



```
(7,
 '0.026*"chat" + 0.019*"duty" + 0.017*"night" + 0.017*"tend" + 0.016*"front" '
 '+ 0.015*"eye" + 0.013*"food" + 0.012*"couple" + 0.010*"interaction" + '
 '0.010*"crazy"'),
(8,
 '0.023*"reach" + 0.023*"environment" + 0.022*"suicide" + 0.021*"system" + '
 '0.020*"hobby" + 0.017*"note" + 0.016*"sector" + 0.016*"lack" + '
 '0.015*"likely" + 0.014*"attempt"'),
(9,
 '0.072*"work" + 0.045*"year" + 0.034*"job" + 0.021*"school" + 0.021*"month" '
 '+ 0.018*"learn" + 0.016*"pay" + 0.014*"study" + 0.013*"course" + '
 '0.011*"private"'),
(10,
 '0.094*"depression" + 0.026*"sleep" + 0.025*"anxiety" + 0.018*"depressed" + '
 '0.017*"diagnose" + 0.015*"home" + 0.015*"employer" + 0.014*"cry" + '
 '0.012*"exercise" + 0.011*"aware"'),
(11,
 '0.090*"doctor" + 0.018*"disorder" + 0.016*"general" + 0.015*"play" + '
 '0.014*"rape" + 0.014*"stand" + 0.013*"judge" + 0.011*"land" + 0.011*"kick" '
 '+ 0.010*"obviously"'),
(12,
 '0.048*"government" + 0.043*"record" + 0.024*"certain" + 0.023*"serve" + '
 '0.018*"phone" + 0.018*"law" + 0.017*"camp" + 0.014*"cure" + 0.013*"solve" + '
 '0.012*"clear"'),
(13,
 '0.046*"interest" + 0.031*"body" + 0.022*"purpose" + 0.016*"field" + '
 '0.016*"join" + 0.016*"aside" + 0.016*"fit" + 0.014*"like" + 0.013*"fun" + '
 '0.012*"soldier"'),
(14,
 '0.064*"delete" + 0.022*"death" + 0.020*"pes_status" + 0.019*"ambulance" + '
 '0.018*"catch" + 0.016*"topic" + 0.016*"progress" + 0.016*"jump" + '
 '0.013*"interesting" + 0.013*"fap"')]
```

Fig. 5. More Topics Discovered by LDA

Conversations about treatment, clinics and costs are seen in a cluster (Figure 6). Terms such as "hospital", "counsellor", "doctor", "medication", "imh", and "therapy" appear in this topic.

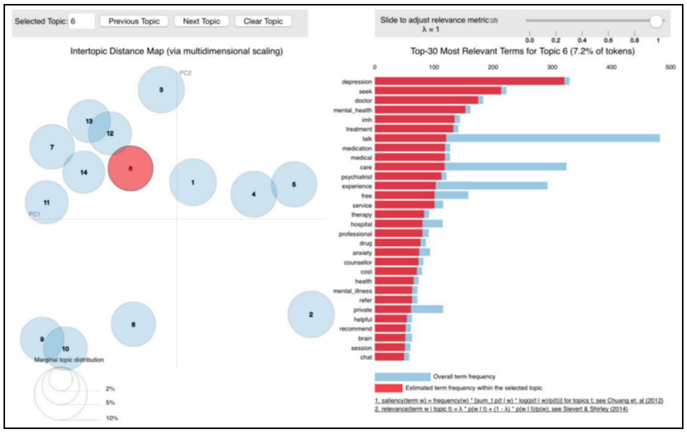


Fig. 6. Results of Topic Modeling. Cluster showing treatment-related terms

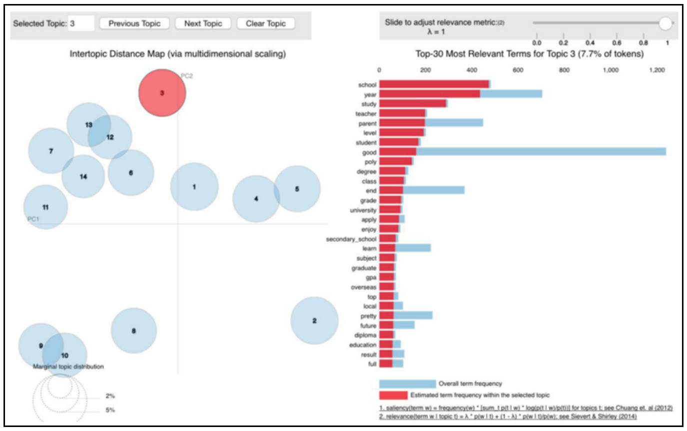


Fig. 7. Results of Topic Modeling. Cluster Showing Terms Related to Schools and Studies.

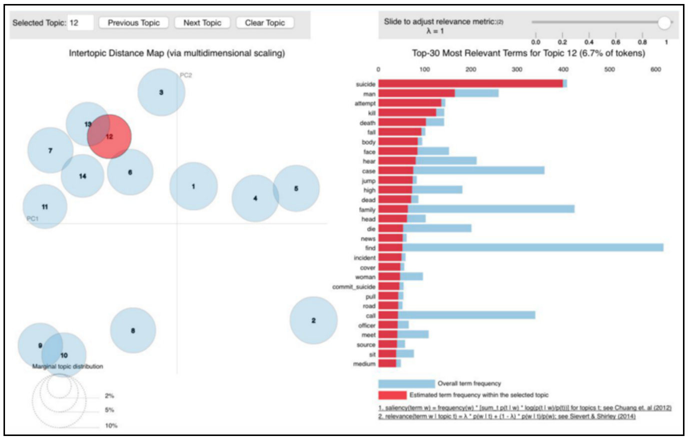
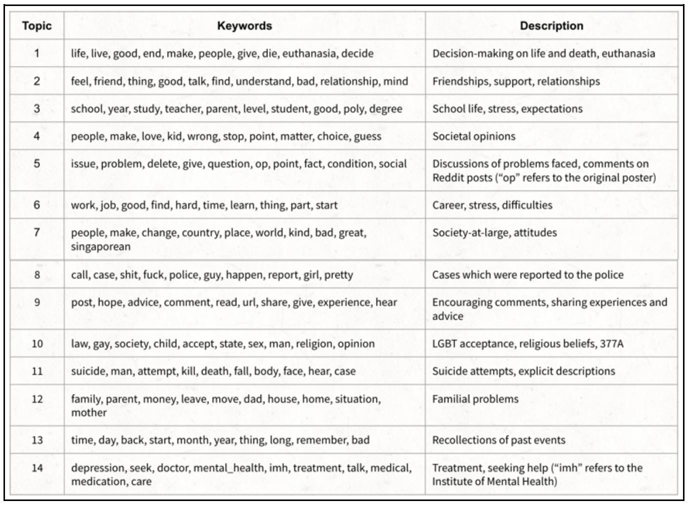


Fig. 8. Results of Topic Modeling. Cluster Showing Terms Related to Suicide Cases and Reports.

Figure 7 shows a topic related to schools and studies. And in Figure 8, we can see another topic with terms related to reports of suicide cases. Below, we contextually explain the topics we discovered using LDA algorithm (Figure 9).



Topic	Keywords	Description
1	life, live, good, end, make, people, give, die, euthanasia, decide	Decision-making on life and death, euthanasia
2	feel, friend, thing, good, talk, find, understand, bad, relationship, mind	Friendships, support, relationships
3	school, year, study, teacher, parent, level, student, good, poly, degree	School life, stress, expectations
4	people, make, love, kid, wrong, stop, point, matter, choice, guess	Societal opinions
5	issue, problem, delete, give, question, op, point, fact, condition, social	Discussions of problems faced, comments on Reddit posts ("op" refers to the original poster)
6	work, job, good, find, hard, time, learn, thing, part, start	Career, stress, difficulties
7	people, make, change, country, place, world, kind, bad, great, singaporean	Society-at-large, attitudes
8	call, case, shit, fuck, police, guy, happen, report, girl, pretty	Cases which were reported to the police
9	post, hope, advice, comment, read, url, share, give, experience, hear	Encouraging comments, sharing experiences and advice
10	law, gay, society, child, accept, state, sex, man, religion, opinion	LGBT acceptance, religious beliefs, 377A
11	suicide, man, attempt, kill, death, fall, body, face, hear, case	Suicide attempts, explicit descriptions
12	family, parent, money, leave, move, dad, house, home, situation, mother	Familial problems
13	time, day, back, start, month, year, thing, long, remember, bad	Recollections of past events
14	depression, seek, doctor, mental_health, imh, treatment, talk, medical, medication, care	Treatment, seeking help ("imh" refers to the Institute of Mental Health)

Fig. 9. Summary of 14 Topics Related to Depression and Suicide.

Our analysis reveals interesting insights into different aspects or themes concerning conversations about depression and suicide. Given a very large amount of textual content, topic modelling was able to quickly summarize and group social conversations about depression and suicide into 14 topics where each topic is uniquely identifiable by a set of keywords.

The discovered topics reveal different contexts in which suicidal thoughts or discussions occur. They may not necessarily be direct causes of suicide acts. However, the insights into the contexts in which conversations about depression and suicide occur can help all of us better understand where to look in order to find troubled individuals in need of desperate help.

IV. Conclusion and Future Directions

Our study demonstrates that a data mining approach allows for efficient and automated ways for detecting suicide in social media. In Singapore, there are two 24-hour suicide hotlines available: 1) Samaritans of Singapore and 2) SAF Hotline (for military personnel). A number of other non-emergency helplines are available such as Youth Line and Tinkle Friend. While these services remain critical in ensuring safety of people, we believe that tapping into digital traces such as public forums can help authorities proactively reach out to those in need of help.

Future directions of this proof-of-concept study include: 1) expanding the current corpus with local languages beyond the Chinese language, 2) working with local organizations to combine digital traces with offline data (e.g. suicide hotlines, counseling), and 3) expanding our analysis to other popular public forums beyond Reddit.

References

[1] Suicide data. (2018 November 11). Retrieved from http://www.who.int/mental_health/prevention/suicide/suicideprevent/en/

[2] Rashith, R. (2018 November 11). "Number of suicides committed by the elderly hits record high as Singapore population ages". Retrieved from https://www.straitstimes.com/singapore/more-than-1-in-3-suicides-committed-by-elderly-as-singapore-population-ages

[3] The Situation in Singapore. (2018 November 11). Retrieved from https://www.sos.org.sg/learn-about-suicide/quick-facts

[4] J. Jashinsky, S. H. Burton, C. L. Hanson, J. West, C. Giraud-Carrier, M. D. Barnes, and T. Argyle. Tracking suicide risk factors through twitter in the us. Crisis, 35(1):51–59, 2014.

[5] Taylor, D. Suicide as a Product of Modernization. Retrieved from https://medium.com/@dvlan/suicide-as-a-product-of-modernization-776e9823cfaf

[6] Fun, P. Lonely 12-Year-Old Girl Sadly Commits Suicide Because She Barely Passed Exams. Retrieved from https://www.worldofbuzz.com/lonely-12-year-old-girl-sadly-commits-suicide-barely-passed-exams/

[7] PRAW: The Python Reddit API Wrapper. https://praw.readthedocs.io/

[8] Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003) Latent Dirichlet Allocation. The Journal of Machine Learning Research, 3, 993-1022.

[9] Gensim: https://radimrehurek.com/gensim/