DSCI 560 Project Proposal

Detect Existence of Suicidality Behaviors in Depression Subreddit

Group name: Team JHP
Group member: Lishi Ji, Jiabao He, Gangyu Pan

1. Abstract

Reddit is a network of communities where people can dive into their interests, hobbies and passions. There are subreddits regarding mental health problems, and people often share experiences, offer support and ask for help on Reddit.Since many people nowadays choose to express their true feelings on Reddit anonymously, it is useful to have an application that monitors posts and prevents people from suffering as much as possible.

The project aims at using natural language processing and data analytics to analyze posts on Reddit. By collecting data from depression and suicidewatch subreddits and using natural language processing and classification models, we would like to predict the probability of users with depression having suicidality behaviors. By listing the posts with detected high probability of existence for suicidality behaviors, we want to ensure the users and posts can receive responses and help as soon as possible.

2. Dataset Description

i. Data Source

We have downloaded the titles and posts contents from depression and suicidewatch subreddits. We save these datasets as csv files locally.

ii. Data Composition

The original depression dataset consists of 100000 rows and 70 columns. This is a sample of depression dataset:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1		all_awarc	allow_liv	author	author_fl	author_fl	author_fl	author_fl	awarders	can_mod_p	contest_m	created_u	domain	full_link	gildings	id
2	0	[{'award_	FALSE	[deleted]				dark	[]	FALSE	FALSE	1.61E+09	self.Suic	https://w	{}	k6y1v7
3	1	[]	FALSE	Sea-Astronomer-7073					[]	FALSE	FALSE	1.61E+09	self.Suic	https://w	{}	k6y1so
4	2	[]	FALSE	InteractionGold1759					[]	FALSE	FALSE	1.61E+09	self.Suic	https://w	{}	k6y0kp
5	3	[]	FALSE	TheOnlyTrueFlame					[]	FALSE	FALSE	1.61E+09	self.Suic	https://w	{}	k6y03z
6	4	[]	TRUE	Notgoodbye75					[]	FALSE	FALSE	1.61E+09	self.Suic	https://w	{}	k6xy4z
7	5	[]	TRUE	ParlorThe					[]	FALSE	FALSE	1.61E+09	self.Suic	https://w	{}	k6xw6i
8	6	[]	FALSE	UrDadsASemenSlurper					[]	FALSE	FALSE	1.61E+09	self.Suic	https://w	{}	k6xuqm
9	7	[]	FALSE	Wavefunctionz					[]	FALSE	FALSE	1.61E+09	self.Suic	https://w	{}	k6xrl2
10	8	[]	FALSE	pewpewmemes					[]	FALSE	FALSE	1.61E+09	self.Suic	https://w	{}	k6xp4u

The original suicidewatch dataset has the same size as the depression dataset.

iii. Data Preparations

We dropped the deleted authors and posts written by deleted authors from the depression dataset, and we left 97849 rows. Also we drop some useless columns and are left with 9 columns at the end.

	author	author_fullname	domain	full_link	num_comments	over_18	score	selftext	title
0	scorpiosofty	t2_88o7w0sl	self.depression	https://www.reddit.com/r/depression/comments/j...	6	False	1	I was in deep clinical depression where nothin...	I've been getting happier recently and I don't...
1	Cult-ImaginOfReddit	t2_3t3qtsxw	self.depression	https://www.reddit.com/r/depression/comments/j...	8	True	1	hiya,so ive gotten back to hurting myself,most...	whats this clear liquid coming from my fresh s...
2	Disappeared777	t2_88o9pbn0	self.depression	https://www.reddit.com/r/depression/comments/j...	0	False	1	I'm desperately wanting to go back and fix my ...	Thinking of the 2019 holidays
3	KNXCV	t2_8n43cvgt	self.depression	https://www.reddit.com/r/depression/comments/j...	2	False	1	Lately, I have been noticing that I get nervou...	Is this social anxiety?
4	brehaorbust	t2_4u3b213w	self.depression	https://www.reddit.com/r/depression/comments/j...	2	False	1	Does anyone else get this feeling? \n\nI' m loo...	I' m panicking about how wrong I feel and how w...
...	...	...	...	...	...	...	...	...	...
99995	AmbassadorLast8446	t2_bax4kf9d	self.depression	https://www.reddit.com/r/depression/comments/n...	4	False	1	I don't get it. My ex best friend and the firs...	Why is she happy?
99996	sheeezy398	t2_g44ca	self.depression	https://www.reddit.com/r/depression/comments/n...	1	False	1	Yet my ex has blocked me on everything and I' m...	I should be happy tomorrow I become a doctor a...
99997	Anubi_Is_Real	t2_1eowe7zh	self.depression	https://www.reddit.com/r/depression/comments/n...	0	False	1	6 days ago I broke up with my gf, but for the ...	I fell empty inside. I don't fell any emotion
99998	TrexCon08	t2_10v3vw57	self.depression	https://www.reddit.com/r/depression/comments/n...	2	False	1	I' ve been kinda struggling with sadness since ...	Hello
99999	OfficerJH	t2_3zkqe0ji	self.depression	https://www.reddit.com/r/depression/comments/n...	1	False	1	Everything feels dull, I can' t distract myself...	I can' t take it anymore

97849 rows × 9 columns

3. Model Description

i. Two Binary Classification Models

With the dataset, the goal is to develop two classification models, one for detecting depression and the other one for detecting suicidality behaviors. We will combine the prediction results from two models to give the final prediction about the existence of suicidality behaviors among the depression.

ii. Deploy Different Models

First, we will split the data into training and test datasets. Then we need to use different techniques for feature extraction which transform each text into a numerical representation in the form of a vector. We plan to use 2-3 methods to build models.
We will begin with the Support Vector Machines model(SVM), since it is a powerful text classification machine learning model and does not need much training data.SVM draws a line or “hyperplane” that divides a space into two subspaces. One subspace contains vectors that belong to a group, and another subspace contains vectors that do not belong to that group.
Then we will also build a model using Convolutional Neural Network (CNN) with Word2Vec for vector representation for words. While dealing with text classification, by varying the size of the kernels and concatenating the outputs, we are able to detect patterns of multiple sizes. Therefore, patterns can be detected no matter where they are located in the sentences.

iii. Model Evaluation and Comparisons

We will evaluate the model performance with accuracy and precision. Then we will end with the model with better performance and use that for later application build up.

4. Project Timeline

Key Milestone	Period	Target Date
Define Phase	Week 3	9/13/2021
Preparation Phase	Week 4 - Week 6	10/1/2021
Modeling Phase	Week 7 - Week 11	11/5/2021
Retrieval Phase	Week 12 - Week 13	11/19/2021
Summary Phase	Week 14	11/26/2021

5. Contribution

Team Member	Content
Lishi Ji	• Model Validation • Web application
Jiabao He	• Feature extraction • Model build up
Gangyu Pan	• Data collection and cleaning • Web application

References

[1] Kim, J., Lee, J., Park, E. *et al.* A deep learning model for detecting mental illness from user content on social media. *Sci Rep* 10, 11846 (2020). https://doi.org/10.1038/s41598-020-68764-y