Connecting Suicidal and Help-Seeking Behaviors on Social Media

Milestone 3

KUSHAL GEVARIA

ADVISOR: PROF. CHRISTOPHER HOMAN

Problem Statement

Differentiate between

- Individuals who have **suicidal thoughts** due to depression, anxiety or mental health issues
- Individuals who just exhibit **general issues** but don't have suicidal thoughts

on social media

Data

Reddit – Provides API to collect data (posts and comments) based on subsections

Subsection related to suicidal thoughts: (SW)

SUICIDEWATCH

Subsections related to general mental/health issues: (GI)

MENTALHEALTH, DEPRESSION, TRAUMA, STOPSELFHARM, SURVIVORSOFABUSE, RAPECOUNSELING, SOCIALANXIETY

Data Insights

Subsection	Total Subscribers in 2015
General Issues (GI)	30,891
Suicide Watch (SW)	4,822

- General Issues subscribers who later subscribed to Suicide Watch = 2,168
- To balance the data set I selected all 2,168 GI -> SW subscribers, plus additional 5,000 GI subscribers
- Total size of the dataset is 7,168 subscribers and their posts

Preprocessing

- **Tokenization** – Split the text into sentences and the sentences into words
- Lowercase the words and remove punctuation
- Remove the words having less than 3 characters
- Remove all the stop words

Example: "I've been feeling depressed on and off for about 2 years, recently there has been more triggers, my anxiety ticks have come back and the depression comes more often (last 2 weeks, everyday). The depression gets worse every time, I've read so many suicide stories, ways to do it etc. but I doubt I would do it but I haven't got that deep yet."

Output: feelings depressed recently anxiety come back depression comes often last everyday depression worse every time read many suicide ways doubt would got deep yet

Preprocessing

- **Lemmatizing** — words in third person are changed to first person and verbs in past and future tenses are changed into present
- **Stemming** — words are reduced to their root form

Example: feelings depressed recently anxiety come back depression comes often last everyday depression worse every time read many suicide ways doubt would got deep yet

Lemmatized output: feeling depressed recently anxiety come back depression come often last everyday depression worse every time read many suicide way doubt will get deep yet

Stemmed output: feel depress recent anxiety come back depress come often last everyday depress worse every time read many suicide way doubt will get deep yet

Feature Extraction

- Convert processed posts (documents) into the bag-of-words

Example:

Processed document: [stress, depress, quit, kill, suicide, die]

Doc2Bow output: [(stress, 5312), (depress, 71886), (quit, 8795), (kill, 9865), (suicide, 16223), (die, 9327)]

Latent Dirichlet Allocation

LDA or latent Dirichlet allocation is a ***generative probabilistic model*** of a collection of composites made up of parts. In terms of topic modeling, the composites are documents and the parts are words and/or phrases (n-grams).

	General Issues	Suicidal Thought
	Topic 0	**Topic 1**
improve	0.271	0.004
kill	0.004	0.565
life	0.360	0.004
progress	0.360	0.004
suicide	0.004	0.425

Table 1: The probability or chance of selecting a particular part when sampling a particular topic

Latent Dirichlet Allocation

Example:

Document 0: [stress, depress, quit, kill, suicide, die]

Document 1: [severe, pain, feel, headache, improve]

Document 2: [kill, suicide, die, pain, quit]

Document 3: [disease, never, worse, bed, sad, medical]

Document 4: [worry, improve, progress, life, therapist]

Document 5: [abuse, social, outcast, fear, unworthy, quit]

	General Issues Topic 0	Suicidal Thought Topic 1
Document 0	0.065	0.935
Document 1	0.924	0.076
Document 2	0.246	0.754
Document 3	0.645	0.355
Document 4	0.924	0.076
Document 5	0.065	0.935

Table 2: The probability or chance of selecting a particular topic when sampling a particular document

Support Vector Machine

Support vector machines are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis.

- Feature used by SVM:

  Matrix of documents v/s words.

  Example:

  [4000 X 2111] => 4000 documents with 2111 unique words

  Final feature list: [stress, depress, happy, quit, kill, suicide ….. therapy]

  Document 0: [1, 1, 0, 1, 1, 1 ….. 0]

Support Vector Machine

Document 0: [1, 0, 0, 1, 1, 0 ….. 1]

Document 1: [0, 1, 1, 0, 1, 0 ….. 0]

Document 2: [1, 0, 0, 0, 1, 0 ….. 1]

Document 3: [1, 1, 0, 0, 1, 0 ….. 1]

Document 4: [0, 1, 1, 0, 0, 0 ….. 0]

Document 5: [1, 0, 0, 0, 1, 0 ….. 1]

Actual / Predicted	**GI->SW (674)**	**GI (1802)**
GI->SW (863)	TP: 470	FP: 393
GI (1613)	FN: 204	TN: 1409

Model accuracy: 75.88%

Using Topic Modeling for SVM

Document 0: [1, 0, 0, 1, 1, 0 ….. 1]

Document 1: [0, 1, 1, 0, 1, 0 ….. 0]

Document 2: [1, 0, 0, 0, 1, 0 ….. 1]

Document 3: [1, 1, 0, 0, 1, 0 ….. 1]

Document 4: [0, 1, 1, 0, 0, 0 ….. 0]

Document 5: [1, 0, 0, 0, 1, 0 ….. 1]

Feature Reduction:

- Extract most frequent words from the topics obtained from topic modeling algorithm

  Example:

  Topic 1 (Suicidal Thought) – [kill, depress, suicide, quit …..]

  Topic 1 (General Issues) – [gone, therapy, headache, improve, progress ….]

Run the SVM model for these selected words instead of all words in the documents

Thank You!

Questions?