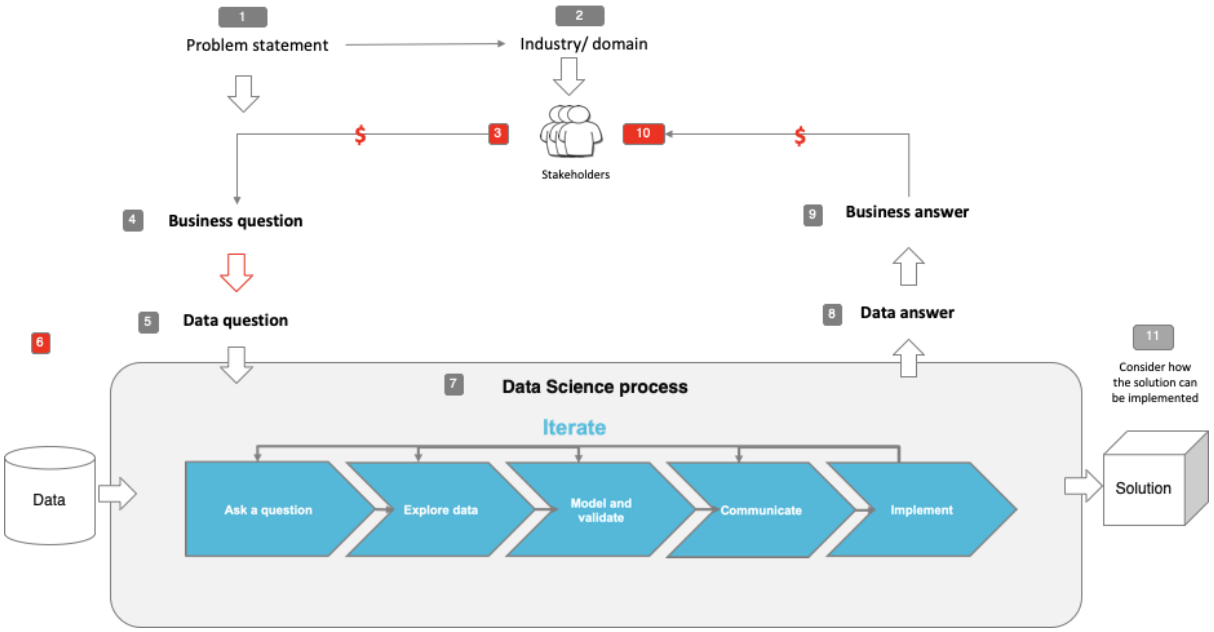
Capstone Project
Document - Johan Ng

Process overview

The following diagram shows the overall end-to-end process for defining, designing and delivering the Capstone project.



Note: The following are the candidate sections of the document. They are presented here for guidance. Questions in each section could be used as possible aspects to cover. Some questions may not be applied to each project. On the other hand, additional information may be needed.

Problem statement

Close to 800 000 people die due to **suicide** every year, which is one person every 40 seconds. Suicide is a global phenomenon and occurs throughout the lifespan

Social Media has become a medium for first hand accounts of persons with mental health conditions due to anonymity. Previous studies have shown that youth are likely to disclose

suicidal thoughts and suicidal risk factors. Thus, there is value in utilizing social media in detecting suicidal thoughts for suicide prevention.

A software start-up specializing in the area of mental health is looking to develop an application to predict suicidal ideation using social media data. The purpose of the app is to act as a

1) Risk monitoring tool
2) Prevention tool
3) Support Community

Thus, the problem statement is ***"Can we use Reddit to detect posts with suicidal ideation?"***

Industry/ domain

Industry:
Social Services Industry

Current State:

- Mostly Academia Research
- Ethical and Privacy Concerns

Stakeholders

Stakeholders:

1) Software Company
    - *Concern*: Utilizing Natural Language Processing to develop a model that is able to accurately detect whether or not a social media post is suicidal

Business question

How do we detect suicidal posts by leveraging on Reddit data and utilize it to recommend interventions to prevent suicide?

Value: To create a support community for people with suicidal thoughts

Data question

What kind of social media data should we collect to build a model for high accuracy of detecting suicidal posts?

Data

Source of Data:
Natural Language Processing Reveals Vulnerable Mental Health Support Groups and Heightened Health Anxiety on Reddit During COVID-19: Observational Study

Volume and Attribute:
Social Media Posts from Jan to Apr (2018 - 2020) from the following subreddits

1) r/depression
2) r/anxiety
3) r/suicidewatch
4) r/alcoholism
5) r/bipolarreddit
6) r/schizophrenia

Ongoing Data:
Data can be collected using Reddit API

Data science process

Data analysis

Data pipeline:
Step 1: Text Preprocessing
The Reddit post was pre-processed using Gensim's preprocessing package and NLTK's lemmatizer package.

The following changes was made to the original Reddit Post

1) Lower case of characters
2) Remove unnecessary whitespace and punctuations
3) Lemmatize word to root form
4) Removing stopwords

Step 2: Topic Modelling
Latent Dirichlet Allocation (LDA) was the statistical modelling technique used for topic modelling

Coherence Score was used as the metric to determine the optimal amount of topics used for the modelling which in this case would be **15 topics**

Hyperparameter tuning of the Alpha and Eta parameters was conducted to determine the best Alpha and Eta values

A dominant topic would then be assigned to each Reddit Post

Based on domain knowledge, the 15 topics are assigned the following topics as follows:

Topic 0: Suicidal Ideation
Topic 1: Anxiety Symptoms
Topic 2: Common Words
Topic 3: Mental Health Conditions
Topic 4: Words used by persons with schizophrenia and bipolar disorder
Topic 5: Strong Negative Words
Topic 6: Coping Mechanisms
Topic 7: Employment
Topic 8: Family
Topic 9: Seeking Advice and Help
Topic 10: Relationships
Topic 11: Sleep
Topic 12: Alcoholism
Topic 13: Academics
Topic 14: Feelings

Step 3: Manual Classification of Suicidal Posts

Based on domain knowledge, **Topic 0, 5, and 9** was assigned to be suicidal posts.

Step 4: Predictive Modelling

GridSearch was done to find out the best vectorizer and machine learning algorithm to be used for the final model.

The main metric used to evaluate the model would be the accuracy score.

Modelling

Main Features:
Cleaned Reddit Posts

Word Vectorizer Used:
TF-IDF Vectoriser package from Sci-Kit Learn

Hyperparameters:

Models Used

Multinomial NB
Standard Scaler + Logistic Regression
Standard Scaler + KNeighbor Classifier
Standard Scaler + Random Forest Classifier

Machine Learning Algorithm	Train Accuracy	Test Accuracy	Precision	Recall
Multinomial Naive Bayes	70%	69%	72%	70%
Logistic Regression	73%	73%	72%	73%
KNeighbors	79%	70%	69%	70%
Random Forest Classifier	96%	76%	75%	76%

Final Model Selected
Random Forest Classifier

Outcomes

Main Findings:

1) Topic Modelling is a good method for unsupervised text clustering in mental health data
2) The topics that are derived from the Topic Modelling is relatively representative of the subreddits
3) The model is able to detect suicidal posts with 76% accuracy

Implementation

Considerations:

1) Data and Privacy concerns should be adhered to and respected
2) Getting more social media data to train the model
3) Using Deep Learning and Neural Networks for modelling

Data answer

Reddit data proved to be useful in the detection of suicidal posts with 76% accuracy. More data from other social media sources needs to be considered to build a more accurate model

Business answer

This project proved that social media data proved to be useful in detection of suicidal ideation in social media users.

Response to stakeholders

There is value in developing an application to detect suicidal ideation among the population and that it is important to take into concern the privacy and ethical concerns in the use of social media data when developing the application

References

Data:
https://zenodo.org/record/3941387#.YG7FdaWA5_R

Low, D. M., Rumker, L., Torous, J., Cecchi, G., Ghosh, S. S., & Talkar, T. (2020). Natural Language Processing Reveals Vulnerable Mental Health Support Groups and Heightened Health Anxiety on Reddit During COVID-19: Observational Study. *Journal of medical Internet research*, *22*(10), e22635.

Notebook:
Capstone Project Part I: CSV Preparation
Capstone Project Part II: Combining Dataset
Capstone Project Part III: Topic Modelling
Capstone Project Part IV: Modelling

Citations:

Doran CM, Kinchin I (2020) Economic and epidemiological impact of youth suicide in countries with the highest human development index. PLoS ONE 15(5): e0232940. https://doi.org/10.1371/journal.pone.0232940

Low, D. M., Rumker, L., Torous, J., Cecchi, G., Ghosh, S. S., & Talkar, T. (2020). Natural Language Processing Reveals Vulnerable Mental Health Support Groups and Heightened

Health Anxiety on Reddit During COVID-19: Observational Study. *Journal of medical Internet research*, *22*(10), e22635.

Shepard DS, Gurewich D, Lwin AK, Reed GA Jr, Silverman MM. Suicide and Suicidal Attempts in the United States: Costs and Policy Implications. Suicide Life Threat Behav. 2016 Jun;46(3):352-62. doi: 10.1111/sltb.12225. Epub 2015 Oct 29. PMID: 26511788; PMCID: PMC5061092.

Roy, A., Nikolitch, K., McGinn, R. *et al.* A machine learning approach predicts future risk to suicidal ideation from social media data. *npj Digit. Med.* 3, 78 (2020). https://doi.org/10.1038/s41746-020-0287-6