# Suicide Detection in Workplaces

# Table of Contents

# Business Goals

What business problems are we trying to solve, and what kind of solutions are we providing

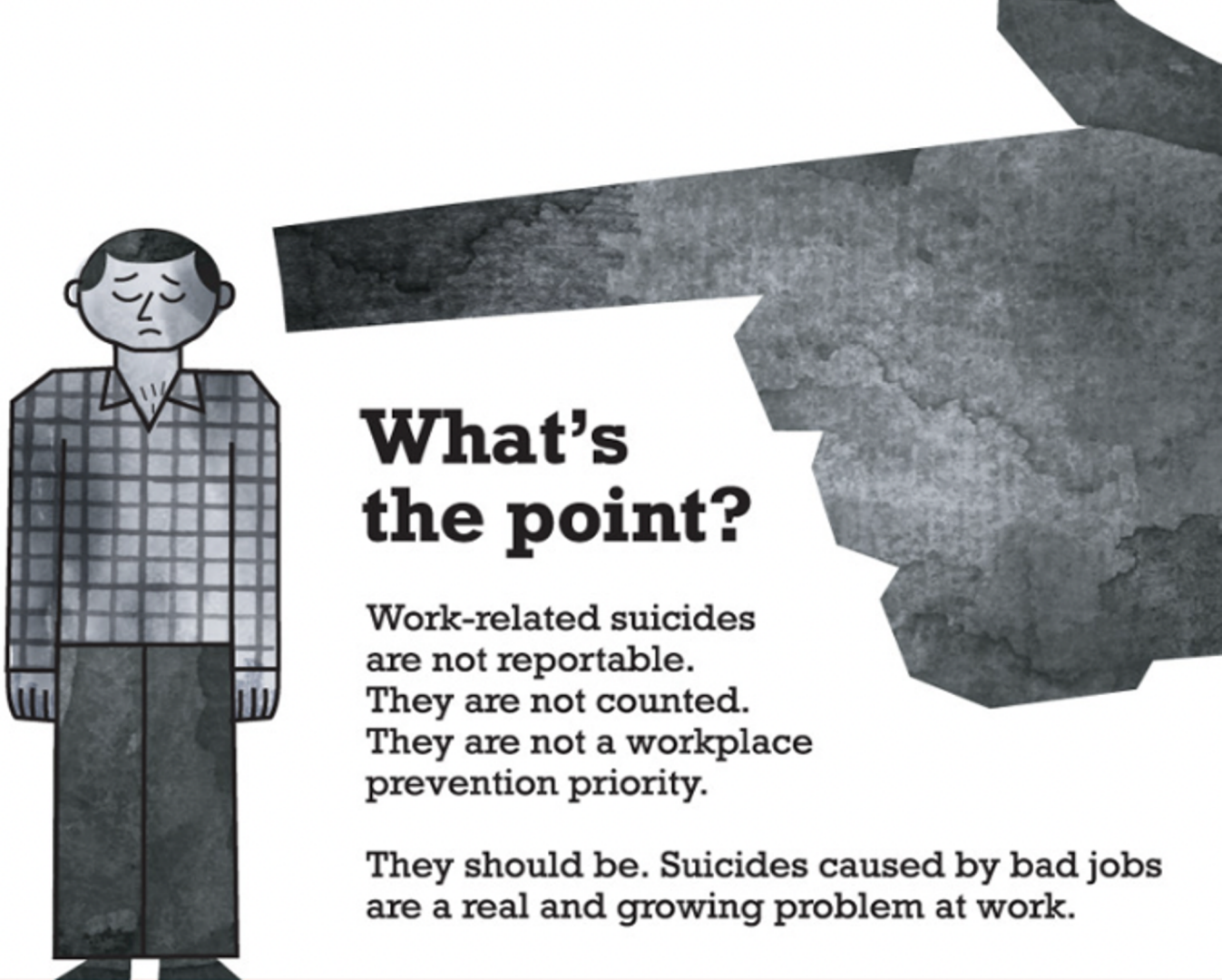



Image Source: hazards magazine

# Problems & Solutions

**Business Problems**

Suicidal employees may cost companies money :

1. Medical Costs (Hospitalization) and/or Compensation(Death)

2. Post-traumatic Stress Syndrome Training for other employees

3. Hurt Brand Images (Potential loss of existing employees, talents who could be onboard or customers)

**Business Solutions**

Goals:

Save costs and improve employees' productivity, leading to increased profits.

Methods:

- Create a suicide prediction tool (Supervised Learning Classification) based on text data
- Identify workers‘ flag behaviors or emotions related to mental health concerns by Topic Modeling

Suicides cases raise at work between 2018 and 2020, according to a report by the *Bureau of Labor Statistics*. Suicides occur most often among people in their working years of 24 to 64 years old.

# Our Data

| | |
|---|---|
| **Source** | The dataset is a collection of posts from "SuicideWatch" and "depression" subreddits of the Reddit platform. The posts are collected using Pushshift API. All posts that were made to "SuicideWatch" from Dec 16, 2008(creation) till Jan 2, 2021, were collected while "depression" posts were collected from Jan 1, 2009, to Jan 2, 2021. |
| **Features** | The dataset displays a text column and a class column (suicide, non-suicide) |
| **Number of Observations** | 232,074 |

| | text | class |
|---|---|---|
| 2 | Ex Wife Threatening SuicideRecently I left my ... | suicide |
| 3 | Am I weird I don't get affected by compliments... | non-suicide |
| 4 | Finally 2020 is almost over... So I can never ... | non-suicide |
| 8 | i need helpjust help me im crying so hard | suicide |
| 9 | I'm so lostHello, my name is Adam (16) and I'v... | suicide |

# Data Preprocessing
# &
# Feature Engineering

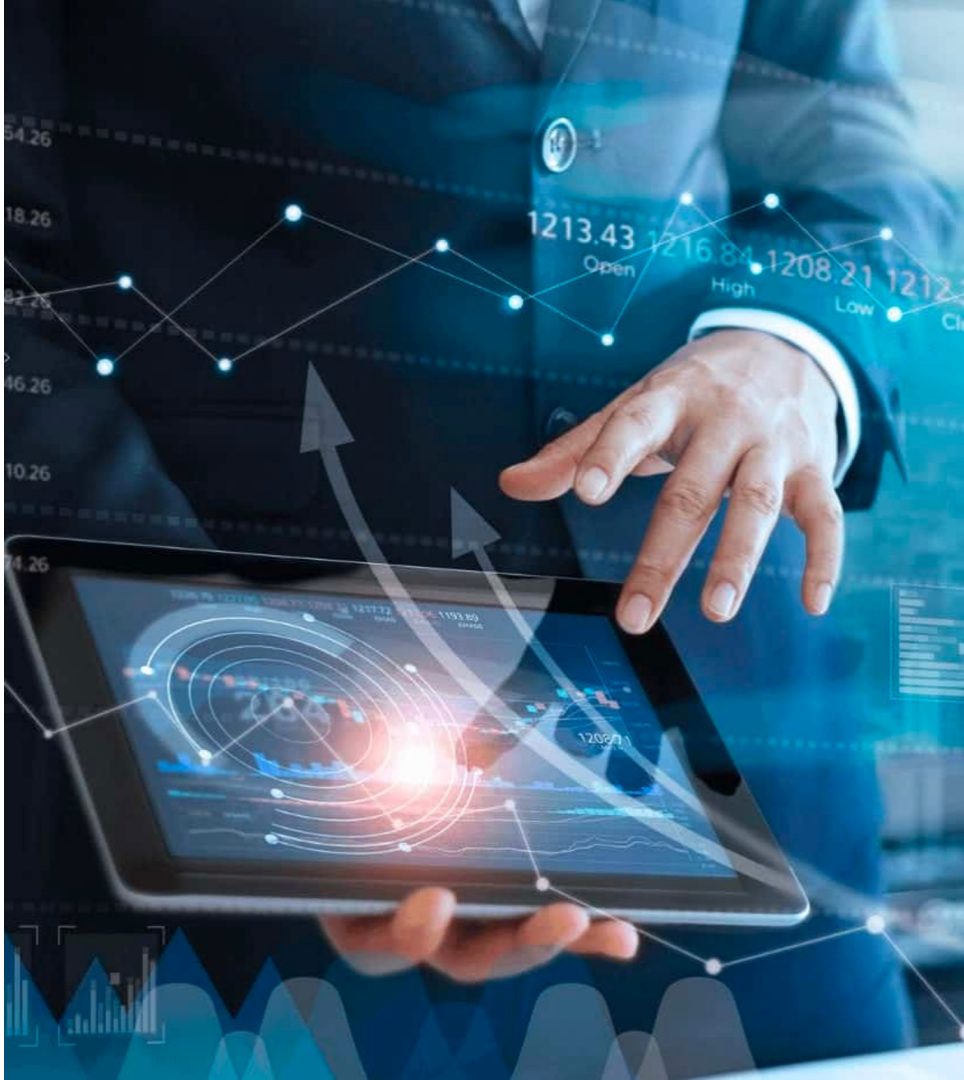



Image Source:
SolveXia

# Data Preprocessing

## Translate emoji to English description

For better interpretation, we replaced emojis with their English names (ex. 😭 → “loudly crying face”)

## Remove stopwords

We removed NLTK (Natural Language Toolkit)'s default list of English stopwords and some unmeaningful punctuations like “.” “,”

## Regex cleaning

Group synonyms or words in the same category (ex. career, intern(ship), position → WORK)

Transform oral language into written language (ex. wanna → want)

## Lemmatization

We used lemmatization over stemming because it converts words to their meaningful base forms.
(ex. ‘Caring’ -> ‘Care’ rather than ‘Car’)

# Topic Modeling

## Bigram

| non-suicide | suicide |
|---|---|
| end life | want die |
| want die | like shit |
| suicidal thoughts | want end |
| want live | want kill |
| want talk | need help |
| … | … |

**Non-suicide**: negative thoughts but want live and talk
**Suicide**: nearly all negative words

## Trigram

| non-suicide | suicide |
|---|---|
| smile face sunglass | face pensive face |
| cool smiling face | downcast face sweat |
| loudly crying face | face tears joy |
| face rolling eyes | face steam nose |
| face tears joy | crying face loudly |
| … | … |

**Non-suicide**: lots of positive emojis
**Suicide**: lots of negative emojis

# Modeling & Performance Evaluation


Image Source: GoodData

# Model Exploration & Architecture

1. Bert Tokenizer
2. Train/Test Sets 50-50 Split
3. Model Exploration
    a. Logistic (baseline)
    b. Random Forest
    c. LightGBM
    d. XGBoost
    e. CatBoost
    f. Bert

**Logistic Regression**

In this algorithm, the probabilities describing the possible outcomes of a single trial are modelled using a logistic function.

**Random Forest**

A meta-estimator that fits a number of decision trees on various sub-samples of datasets and uses average to improve the predictive accuracy of the model.

**LightGBM**

A fast, distributed, high-performance gradient boosting framework based on decision tree algorithm, used for ranking, classification, etc.

**XGBoost**

A scalable, distributed gradient-boosted decision tree (GBDT) machine learning library that provides parallel tree boosting.

**CatBoost**

It provides a gradient boosting framework which among other features attempts to solve for Categorical features using a permutation driven alternative compared to the classical algorithm.

# Model Exploration - Bert

## Bert

which stands for Bidirectional Encoder Representations from Transformers

01

### BertTokenizer

Use BertTokenizer from pre-trained and set truncation, MAX_LEN = 50, and batch_size = 32

02

### Bert Classification

Use BertForSequenceClassification from pre-trained

03

### Optimizer & Parameters

Optimizer: AdamW(lr=2e-5, correct_bias=False)

- AdamW: a variant of the optimizer Adam that has an improved implementation of weight decay which is a form of regularization to lower the chance of overfitting

Parameters (109,483,778 trainable parameters)

- weight_dacay_rate

04

### Evaluation Function

Accuracy & Loss

# Model Performance

| L | R | L | X | C | B |
|---|---|---|---|---|---|
| **71.8%** | **77.1%** | **82.2%** | **85.7%** | **85.7%** | **97.4%** |
| Logistic (Baseline) | Random Forest | LightGBM | XGBoost | CatBoost | **Bert** |
| Accuracy: 0.7177 | Accuracy: 0.7707 | Accuracy: 0.8223 | Accuracy: 0.8574 | Accuracy: 0.8573 | Accuracy: 0.9740 |
| Average Runtime: 7.21 s | Average Runtime: 1 min 31 s | Average Runtime: 8.45 s | Average Runtime: 1 min 40 s | Average Runtime: 1 min 45 s | Average Runtime: 28 min 23 s |

# Business Recommendations

Image Source:
Make a Difference

# Implementation Roadmap

1. **Suicide Prediction Tool Use**
   - Mine online posts (social media, blogs, etc.) or messages from workplace communication tools for words linked to suicide
   - Identify flag behaviors or emotions to alert the companies of their employees' mental health issues.

2. **Workplace Suicide Prevention Program**

   Suicide prediction tool can be incorporated to the Workplace Suicide Prevention Program in companies.
   - Once a flagging suicidal reason is identified by the tool, the program will promote help-seeking and help-giving to the targeted employee at the workplace.

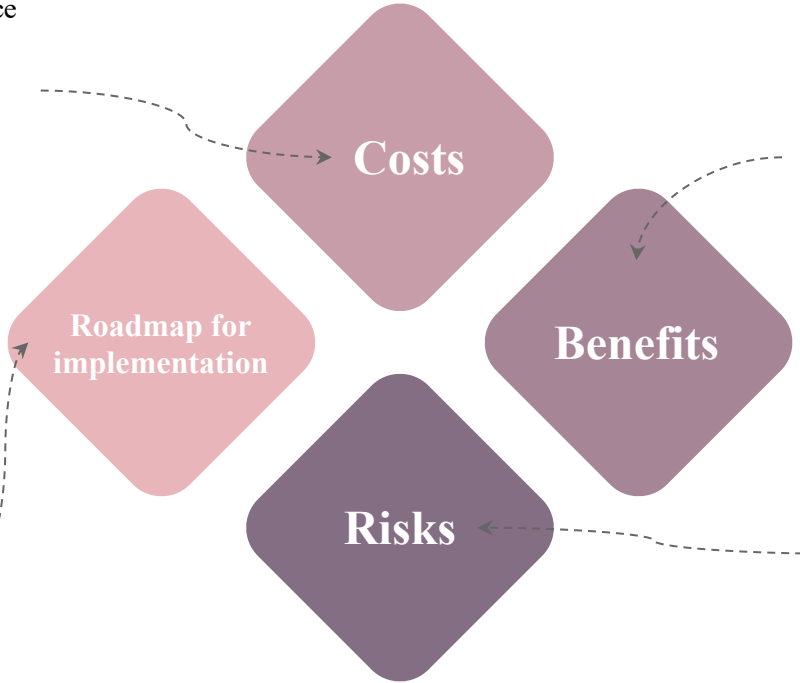
3. **Improvement & Concern**
   - Our tool is created based on posts on Reddit that are not only posted by employed people. We would like to collect more employee data in the future and rebuild our tool.
   - Using online data for suicide prediction is still subject to oversight and review to ensure its effectiveness, safety, and ethical permissibility.

# Return on Investment

Image Source:
Talk space

# Return on Investment

1. Enough data on the cost of inpatient psychiatric care for suicidality, police and EMS response, crisis institutes, and so on.
2. Allocation of human resources that includes data cleaning and time for model making.




1. Based on existing research literature, we estimate that each year of investment in the suicide prediction tool will at least prevent 3,600 suicide attempts and 140 deaths over the next 28 years. This leads to a reduce of 0.13% in the number of suicide attempts and deaths.
2. We project the financial benefits of stoping these attempts and deaths to be $1,100 per dollar invested in the tool. These benefits include savings in medical costs and increased profits.

Implement suicide prediction tool in the company and incorporate it to Workplace Suicide Prevention Program.

- Once high risk people are identified, provide follow-ups or cognitive behavioral therapy.

The correlations described by the results are ONLY applicable to the particular data sets used within the specified location.

# Conclusion

Image Source:
Tesis Master

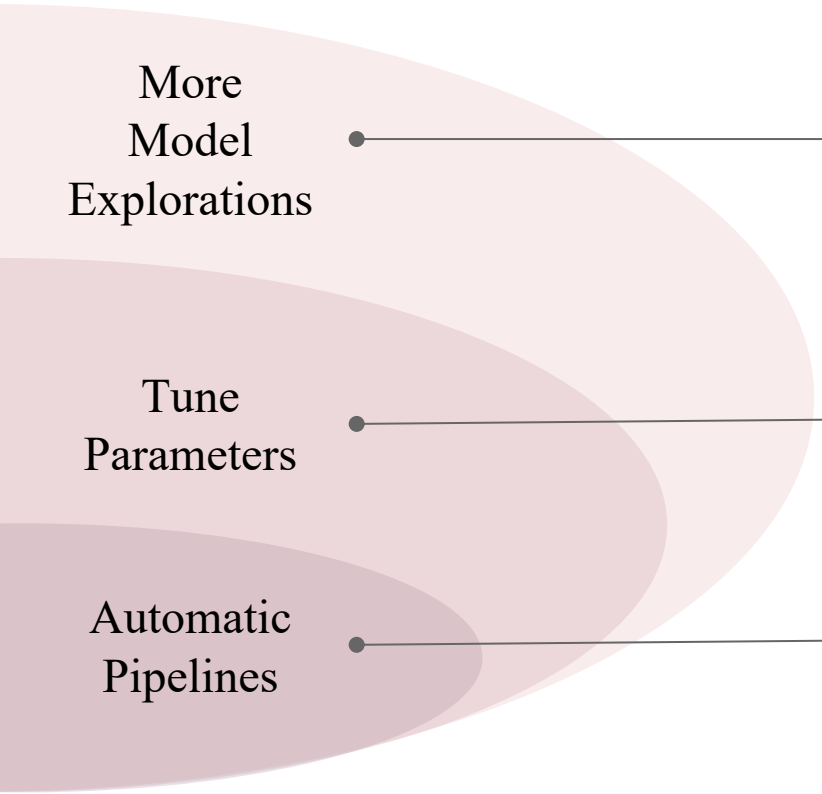
# Conclusion

- Advanced NLP techniques can create business opportunities and identify potential suicidal actions at workplaces.
- Bert Model gives the highest classification accuracy of 97.4%, but it also has the longest runtime of 28 min 23 s. This could be implemented into businesses' mental health program or a suicide detection product that can be widely used.
- By looking at the result of model performance, we can promote help-seeking and reduce the incidence of suicide and suggest suicide-prevention campaigns.
- Suicide prediction can result in significant potential cost savings as a result of fewer suicide deaths and reduced life years lost.

# Future Improvement

Image Source:
accounting web

# Future Improvement



More Model Explorations

1

For efficiency, we would like to try fasttext, DistilBert, FastBert and so on.
For accuracy, we would like to try more advanced models which are improved based on Bert such as RoBerta, DeBerta-V3.
Finally, we would like to find the balance (Trade-offs) between efficiency and accuracy.

Tune Parameters

2

We did not have enough time to tune parameters. Hence, we would like to tune parameters to control running time, limit overfitting, and improve accuracy in the future.

Automatic Pipelines

3

Functionalize things, automate workflows and create pipelines to deal with streaming data.

# Reference


https://analyticsindiamag.com/7-types-classification-algorithms/

https://towardsdatascience.com/lightgbm-vs-xgboost-which-algorithm-win-the-race-1ff7dd4917d

https://www.nvidia.com/en-us/glossary/data-science/xgboost/

https://en.wikipedia.org/wiki/Catboost

https://healthitanalytics.com/news/suicide-risk-prediction-models-prove-cost-effective-in-healthcare

https://www.rand.org/pubs/periodicals/health-quarterly/issues/v5/n2/09.html