securly://

Understanding NLP

Or how Securly (almost)
aced the Turing test

May 2020

Abstract

This paper provides a transparent overview of the engineering that allows Securly to detect students in danger with an almost 95% accuracy. Securly deploys complex machine learning algorithms that analyze data against carefully curated data sets. Securly's Natural Language Process (NLP) engines are put through rigorous training and multiple levels of data analysis that train them to think like humans when detecting grief, depression, bullying, self-harm, and suicidal thoughts in kids. Given the impact this has on the ability to save lives, Securly's NLP engines are being trained to ace the Turing Test - an industry benchmark that defines the viability of any AI engine.

Table of contents

Introduction

There has always existed a level of fragility to human emotions, which is why experiences like depression, sorrow, even suicide ideation, while unfortunate, are nothing new to the human condition. But when it comes to school-aged kids still in the process of developing emotionally, the implications of such emotions can at times be fatal.

Technology is learning to be empathetic toward human emotion by recognizing signs of trouble which might otherwise go overlooked. AI enables Securly to recognize kids who are suffering and then intervene to make a positive difference. Back in 1950 when the Turing Test was proposed, the evolution of machines that could ace the test and imitate human thinking seemed fantastic. Today it is possible. Currently, Securly's NLP engines are able to recognize signs of distress in students with a ~ 95% rate of accuracy. When woven into the fabric of K-12 tech, this can directly be measured in the number of students saved.

What is the turing test?

The Turing Test, proposed by famed computer scientist and mathematician Alan Turing, tests the capability of a machine to think like a human, an important marker of the viability of any AI engine. When put through the Turing Test, the machine should be able to convince a judge (usually a human) that the machine is, in fact, a human. If a machine passes the Turing Test, then the machine has proven itself to be as intelligent as and having the ability to think like a human being.

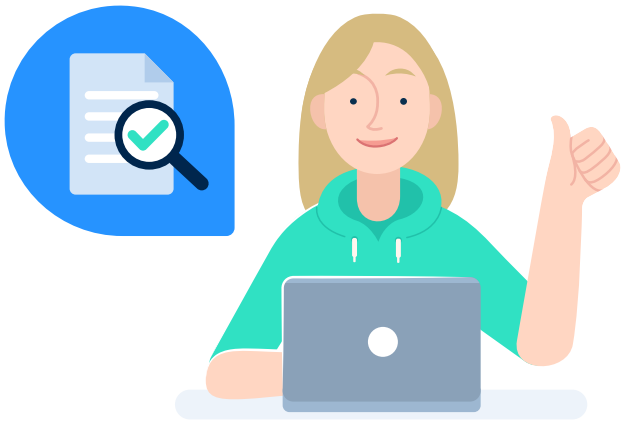
Given the sensitivity of information Securly's AI engines process and the impact it has on student lives, our engines are trained in a way that allows them the ability to imitate human thinking with an almost 95% accuracy. When our 24 teams and NLP engines analyze the same datasets, the engines interpret the data in the same way the human team does almost 95% of the times. This has huge implications not just as a reliable source of detection, but also for reducing the reaction time for parents/ schools/ first-responders when tragedy seems to be imminent.

How does NLP work at Securly?

The process that seems as simple as

> input post >> analyze sentiment >> issue alert/pass through

is, in fact, a long computational process that involves complex machine learning algorithms that analyze incoming data against carefully curated datasets.



1. Data Pre-processing

All raw data coming into the system needs to be pre-processed to prepare it for our sentiment analysis engine. This pre-processing of data scraps off the noise in it such that there are no words or characters, deemed unnecessary for sentiment analysis, left over in it.

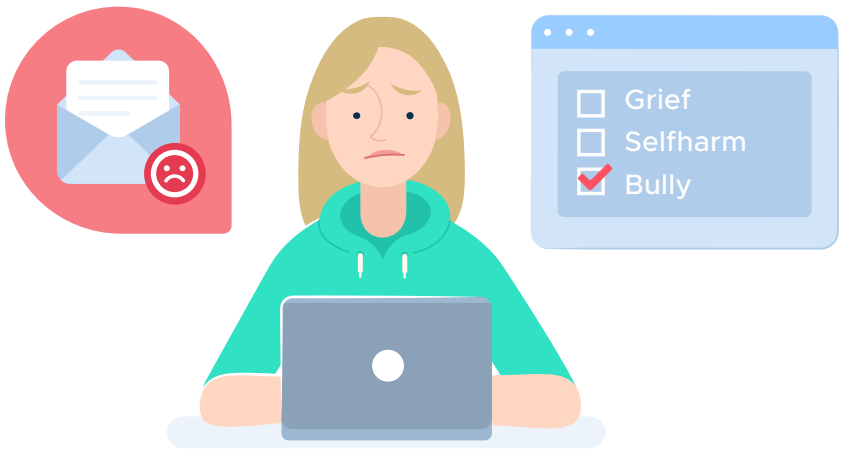
a. Removing stop-words: As a first step, our algorithm removes all commonly occurring words such as ‘the’, ‘a’, ‘an’ etc. These words do not add special meaning to the text and it can be understood just as well without it.

b. Customized Securly pre-processing: Considering the frequency of spam text and certain other elements of electronic communication we introduce a second round of processing where we remove elements that are not part of the actual "conversation" that we are interested in understanding. This includes timestamps, attachments, meta tags, URLs, Twitter handles etc.

c. Lemmatization: This is a very important part of the process which helps us derive the base word of every word we encounter. A standard NLP library that gives the base word for every word possible is used for this purpose. Lemmatization takes into account the morphological information of a word and provides its base word. E.g. car, cars, car's the base is car; for am, are, it is be. Lemmatization is more than just looking at the stem of a word and is very important when it comes to analyzing the sentiment behind a sentence.

2. Classification

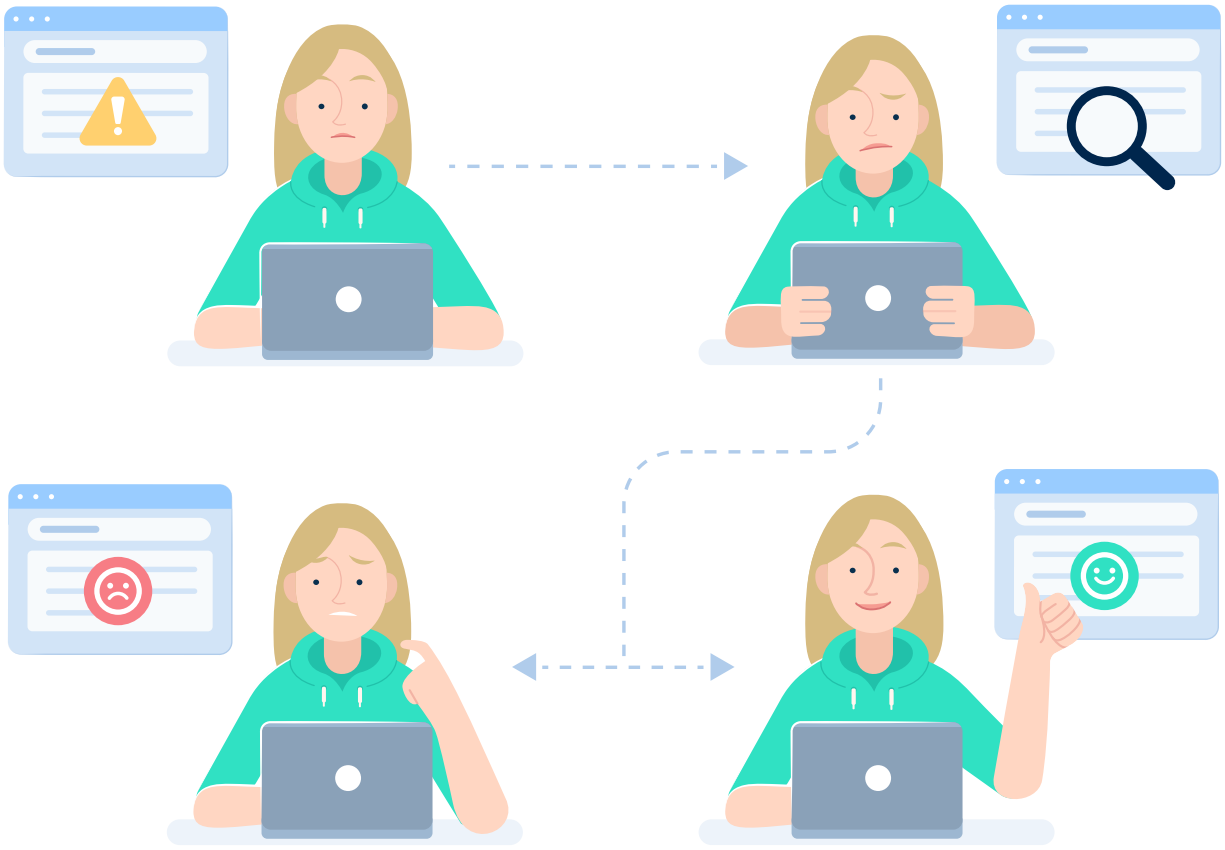
This pre-processed data is then sent to the sentiment analysis engine for classification. The data is classified as either clean, bully or grief. Identifying a text as bullying or grief requires rigorous training of the engine. (We will go into the depths of how the engine is trained in subsequent sections.) If a text is identified as clean, it is checked against a special dictionary for filth. No matter what you throw at it, our engine can accurately predict more than 80% of the time if the text is grief, bully, or clean.



3. Alert generation

All grief and bully instances trigger an alert to the respective school admins and parents. It is also shared with our 24 teams who analyze it individually for criticality and contact the school, parents or police as required.

The flagged text is also displayed as flagged activity to the school admin (and counselors or teachers if delegated by the admin) on their Securly admin portal and is available at all times for reference. It is possible to download student specific and time-range based activity reports as well if required to share with principals, parents, school boards etc.



Training the engines

The accuracy of Securly's NLP and sentiment analysis engines is dependent upon their training, and so we lay great stress to ensure that they are trained above industry standards. The entire pre-processing of data is done during the training phase for the training datasets as well.

1. Data generation

The Securly datasets are completely driven by the real-life data generated by its filtering products. We maintain separate datasets for each of the three classes – grief, bully and clean. The words and phrases in these datasets have a 1:1 mapping with the labels. This dataset is regularly updated with more words and phrases to ensure greater accuracy of classification.

The dataset for filth is a dictionary and is not used in the training of the engines as it does not require any machine learning algorithms for identification of filth during analysis.

2. Vectorization using N gram range

To help us understand data better it is important to break it down into smaller parts or vectors. This helps analyze a given sentence more accurately. To do this we use the trigram range method. For example, the text – I am going out – is broken down into groups of one word, two words, and three words. These N-grams are called feature.

In our example, the features for 'I am going out' will be:
One word: I + am + going + out = 4 features
Two words: I am + am going + going out = 3 features
Three words: I am going + am going out = 2 features
Which leads to a total of 9 features for this particular sentence.

Such feature sets are created for every text in the dataset and give us an exhaustive dictionary of unique features to work with.

Thereafter, a scoring mechanism gives us a score for each of the features. If a feature occurs too many times it is considered commonplace and its score reduces. The scoring mechanism is important in identifying the total score of a sentence and classifying it as clean, bully or grief.

We also use the chi-squared statistics method to help identify the best features to be used in an engine. This is necessary considering the volume of the dataset, which generates a large number of features. Using all the features without prioritization can impact the efficiency and accuracy of the engine.

3. Classifier/ Classification

Here we use supervised Machine Learning algorithms such as logistic regression among other things to help us draw boundaries around features to determine their class. Considering that every sentence the engine analyzes is a group of features, it is important to lay down these boundaries to help us determine what class most of the sentence falls into. The thickness of these boundaries is also determined by these algorithms and helps us deal with grey areas efficiently.

4. Cross validation

A process of cross-validation is used to separate out three sets of 10% of the datasets, which are then used to test and train the engine. Each set is put through all the parameters during testing and the results are compared to identify the best engine.

A hold-out set is also identified and set aside before the other sets are put through training. The best engine is then put to test against this hold-out set. The hold-out set remains constant and is used as the baseline against which every new engine created has to pass before it is released for use.

The Future

Working up from this foundation for NLP that currently processes shorter text (1 sentence) within 1.87 seconds and longer text (approx. 4KB) within 2.21 seconds, we are working on incorporating the neural networks method as well. Research is underway on this aspect and will help us achieve onwards of 90% accuracy when released.

Also in progress is our Correlations project, which analyzes and assigns a rating to individual students that corresponds with the seriousness and urgency of their online activities. The lowest rating reflects harmless and everyday content, the highest indicates an imminent threat to a person's life. This will give schools a broader view of a student's suspicious online activity, and help them determine when an incident needs intervention.

In addition, we are also constantly updating our hold-out set; and working on a spell-correction algorithm that will help us auto-correct misspelled words, mark them as distinct features and give us better insight into such text. The difference between a situation missed and a student saved can be as small as a typo. As the Student Safety Company, Securly is dedicated to filling in those cracks.

Conclusion

There are many movies where computers become too smart and attempt to eradicate humanity. These cautionary tales generate income for the storyteller but simply do not reflect the reality of the advancements being made in technology.

The reality is that technology like Natural Language Processing and Sentiment Analysis are key to providing Securly and schools an extended awareness of those times when kids may feel alone or, worse, utterly hopeless. In situations where time is of the essence, technological advances allow us to be more aware of and more connected with those in need of help. And if further advancements result in additional lives saved, then rest assured those advancements will be made, and the Turing Test aced.

securly://

sales@securly.com

1-855-SECURLY

www.securly.com