**SOLUTION DEVELOPMENT AND IMPLEMENTATION REPORT**

Toxic Comment Classification

### 

### 

### 

### 

### **PREPARED FOR**

Nethra Sambamoorthi, PhD

PREDICT-498 Section 56

Northwestern University, Winter 2018

### **PREPARED BY**

**Team 4: Green**

Ellis, Joseph

Herena, Ricardo

Law, David

Mesen, Luis

# Table of Contents

[Table of Contents 1](#_Toc506569891)

[Abstract 3](#_Toc506569892)

[Introduction 4](#_Toc506569893)

[Company Profile 5](#_Toc506569894)

[Business Problem and Research Questions 5](#_Toc506569895)

[Online Toxicity Issue Tree 7](#_Toc506569896)

[Hypotheses 7](#_Toc506569897)

[Train and Test Data 8](#_Toc506569898)

[Data QA Report 9](#_Toc506569899)

[Variable List and Data Dictionary 12](#_Toc506569900)

[Data Preparation 13](#_Toc506569901)

[Variable Reduction 13](#_Toc506569902)

[Measurements & Metrics 15](#_Toc506569903)

[Methodology 16](#_Toc506569904)

[Computational Methods 17](#_Toc506569905)

[Output Summary 20](#_Toc506569906)

[Analytics Project Plan 22](#_Toc506569907)

[Project Management Approach 22](#_Toc506569908)

[Project Deliverables 23](#_Toc506569909)

[Project Technical Requirements 24](#_Toc506569910)

[Project Constraints 24](#_Toc506569911)

[Project Assumptions 24](#_Toc506569912)

[Project Risks 25](#_Toc506569913)

[Gantt Chart 26](#_Toc506569914)

[Cost 27](#_Toc506569915)

[Resources 28](#_Toc506569916)

[References 30](#_Toc506569917)

# Abstract

Online harassment in the form of toxic comments is a major impediment to companies that are looking to engage with their customers online. This form of online harassment can be difficult to identify, costly to regulate and can lead to significant emotional distress that ranges from hurt feelings to serious violations of personal rights.

In this paper, we introduce an applied analytic methodology to aid in the detection of this type of harassing behavior. Using a large corpus of labeled data, we design a proof-of-concept pipeline that can pre-processes large volumes of text data and apply machine learning techniques to it in order to identify online harassment accurately and quickly via an interactive dashboard. This can enable organizations to quickly identify harassment and take action, thereby improving customer experience and user engagement on that organization’s online platforms. This document focuses on developing a project plan to analyze, test and implement this analytic method for identifying toxic comments.

We propose a solution that can be used by any company looking to improve its online presence. In conjunction with our business partners, we have identified several datasets, business problems and hypotheses, metrics and methodologies to deploy an effective toxic comment classification application.

# Introduction

In 2017 the digital world saw a marked difference in the way companies and individuals addressed the subject of online harassment. Not only did major technology companies like Twitter, Facebook and Google make significant changes in the way their platforms dealt with online harassment but there was also more attention on this topic from policymakers. Rep. Katherine Clark (D-MA), for example, introduced a bill to the U.S. House of Representatives intended to protect the safety of people experiencing harassment, especially threats of extortion and violence. The Pew Research Center surveyed more than 4,000 U.S. adults and found that approximately “41% of American have been personally subjected to harassing behavior online” (Duggan, 2017). Another study conducted by Data & Society and the Center for Innovative Public Health Research found that 36% of internet users experienced direct forms of harassment (Young, 2017). The negative effects of online harassment can range from embarrassing to physically threatening. Unfortunately, this type of harassment can be difficult for humans to identify or quantify the severity which exists on a spectrum and can easily be overlooked, especially considering the amount of data needing to be analyzed.

For technology companies cultivating a positive experience for their customers, this kind of online harassment can mean less customer engagement and negative customer sentiment. Twitter released significant updates to its platform in March to reduce abusive content and let users take back control of their safety and security using a revamped reporting functionality. Unfortunately, online harassment in the form of toxic comments can plague almost any business with a website, not just large tech companies. Companies require a method to identify and stop the abuse quickly and effectively. For this reason, the Conversation AI team, a research initiative founded by Jigsaw and Google are working on tools to improve online conversation. Team 4: Green intends to build on this effort to help improve detection of toxic comments using predictive analytics methods.

# Company Profile

Team 4: Green is a small, agile predictive analytics company of 4 members. These members bring their mix of talents and expertise in the field of predictive analytics to solve real-world problems with innovative forward-looking solutions. Team 4: Green has a passion for improving people’s experiences online and bringing the humanity back into digital interactions.

|  |  |  |
| --- | --- | --- |
| **Team: 4 Green** | | |
| **Member** | **Title** | **Role** |
| Ricardo Herena | Team Lead | Communications and Engagement Management |
| Joseph Ellis | Analyst | Programming Languages |
| David Law | Analyst | Analytics Modeling |
| Luis Mesen | Analyst | Report Writing |

# Business Problem and Research Questions

While online discussions are a great way for companies to engage their user base and build communities, they may also be plagued by users who harass others using toxic comments. This harassment leads to reduced engagement when toxic comments are allowed to proliferate. Moderation of conversations via human monitoring is costly and limited, especially when conversations are occurring at high volume and velocity. On average, it would take nearly 30,000 moderators inspecting 12 comments a minute to review the number of messages posted on Twitter every minute. An analytical solution that can autonomously identify threat of abuse and harassment in online comments by analyzing patterns and keywords would be greatly beneficial. This solution seeks patterns in the text that are indicative of toxicity, such as use of profanity or excessive punctuation. Identifying toxic comments such as these in an accurate and timely fashion allow organizations to better enforce civil dialog on their online platforms and ultimately cultivate a better online experience for their customers. The business problems and research questions are as follows:

**Business Problems**

* A small subset of hostile users’ harassing behavior reduces customer experience for the wider online community.
* Toxic comments are challenging and time-consuming for online companies to classify and correct manually.
* Online harassment policies are unenforceable and therefore enable toxic comments to run rampant.

**Research Questions**

* Can we improve bad customer behavior by better enforcing the organization’s harassment policies?
* Can we improve customer experience by effectively identifying toxic comments?
* Is there a consistent definition of toxicity in online comments?
* How do different classes of toxicity compare in terms of impact and detectability?
* Are there key words or lexical patterns that are predictive of toxic comments?
* How should the model be used to most effectively reduce online harassment?
  + Should the model be used as a preventative measure? (Users are not allowed to contribute content that does not pass the test)
  + Should the model be used as a corrective measure? (Content is reviewed using the model)
* How should the solution be designed to maximize customer buy-in and implementation?
* What metrics should be defined for the customer to demonstrate that the solution is useful?

# Online Toxicity Issue Tree

In order to create a mutually exclusive, collectively exhaustive set of hypotheses and business issues an issue tree has been constructed in fig. 1. This framework will be used while engaging the business in order to make sure all avenues for resolution of the business problem can be evaluated.

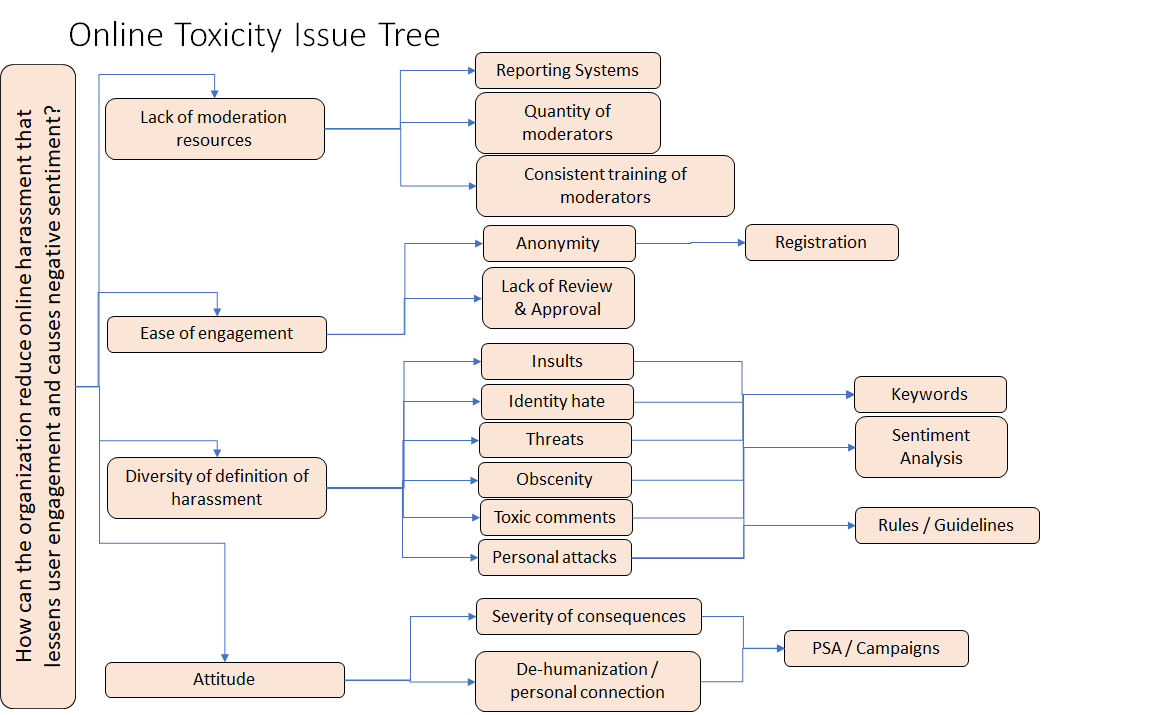


Figure 1 - Online Toxicity Issue Tree

# Hypotheses

Team 4: Green has formulated several hypotheses to answer the research questions above:

* Toxicity “levels” can be created in a way such that a higher index correlates with a higher probability of toxicity
* Toxicity exists in a spectrum of severity; thus, definitions of toxicity will vary with each individual opinion, however there is a threshold above which a comment is universally rated as toxic
* Toxic words and phrases form a distinguishable cluster in the vector space of the corpus
* Certain keywords and phrases are strongly predictive indicators of the severity and class of toxicity
* Despite complications from polysemy and synonymy, it is possible to train a complex natural language model that classifies toxic comments more accurately than a random model
* Reduced toxicity is correlated to increased user activity and engagement
* Comment toxicity is predictable based on text data such that a standard online harassment policy can be enforced
* Consistent application of a standard online harassment policy as defined by each individual organization can reduce harassing customer behavior

# Train and Test Data

The data provided consists of a large number of Wikipedia comments which have been labeled by human raters for toxic behavior. The data is divided into a test and training set, consisting of a total of 312,735 records.

The test set consists of 153,164 records. Each record consists of two columns: ID and Comment\_Text. The ID is a unique identifier for each record and the Comment\_Text consists of a string of text sourced from Wikipedia comments. The training set consists of 159,571 records. Each record consists of seven columns: ID, Comment\_Text, Toxic, Severe\_Toxic, Obscene, Threat, Insult and Identity Hate. The columns Toxic, Severe\_Toxic, Obscene, Threat, Insult and Identity\_Hate consist of binary responses of 1 and 0, with a single comment potentially having multiple classifications. These comments have been evaluated and classified as toxic comments by human raters, with 143,346 of these records containing no classification.

# Data QA Report

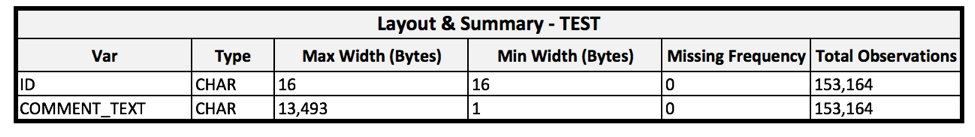
In order to create a standard pre-processing pipeline, a data QA report must be completed for each corpus provided for analysis. Conducting a thorough data quality analysis on unstructured data is a complex, multi-step process with several available methodologies. Ensuring that the proper techniques are utilized in the data’s preparation is paramount to the success of the modeling process as the resulting structure provides the foundation for the effectiveness of classification. Any corpus provided for analysis will be checked for the following issues:

* **Completeness:** Are the number of observations received is the same as the number of observations sent?
* **Segmentation:** Was the data sample segmented in such a way that would degrade the predictive ability of the model? (e.g. only comments from males or females were included in the sample)
* **Sampling Bias:** Was the sample is collected in such a way that all members of the intended population are equally represented? (e.g. toxic comments in the sample are roughly proportional to toxic comments in the population)
* **Sample Weighting:** If there exists sampling bias, was any sample weighting performed to correct for that bias and was it done in a statistically rigorous manner?

The data utilized for the toxic comment classification process will be provided in comma separated values (CSV) format and will consist of text comments composed of strings with no metadata provided. Additionally, binary indicators (1 or 0) are provided based on six types of classification. Our ability to ensure data quality is directly related to our ability to correctly tokenize and tag our data and encompasses the following components:

* **Text Encoding:** Depending on the software utilized for Text Mining, the required format will vary and could include ASCII or UTF-8.
* **Data Parsing:** This includes our ability to properly substring, handle delimiters and compress erroneous data. The use of TreeTagger is under consideration, which utilizes PERL for parsing.
* **Part of Speech (POS) Tagging:** POS Tagging encompasses our ability to correctly identify the semantic parts of the English language. There are currently several English based Tagsets available for this process. The selection of the appropriate Tagset will be critical to ensuring data quality. In addition to the selection of the Tagset itself, there are several learning-based techniques available, including: Maximum Entropy Markov Model (MEMM), Conditional Random Field (CRF) and Hidden Markov Model (HMM). Our selection of an appropriate technique will be critical to ensuring data quality. Additional components of effective tokenizing/tagging which will need to be considered include the use of: stemming, lemmatization and stop words.

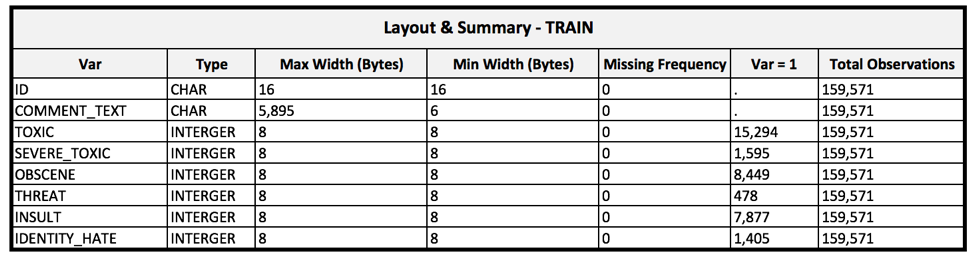
The following tables provide summary information for the test set, including: file layout, missing frequencies, total observations and the first 10 observations from the set itself. Any dataset provided will be required to conform to the specifications below:



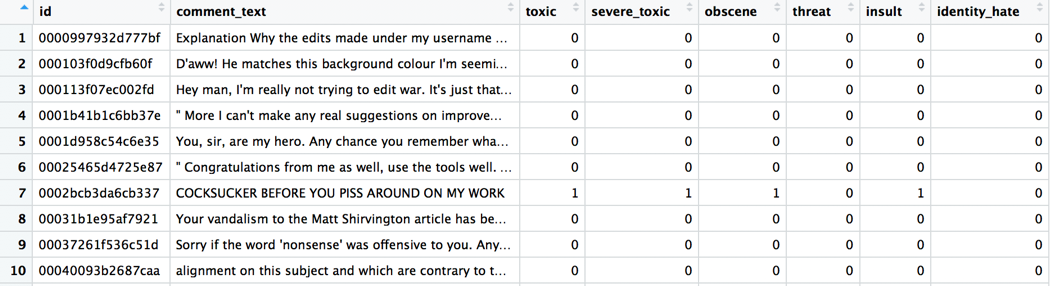
Sample test data:



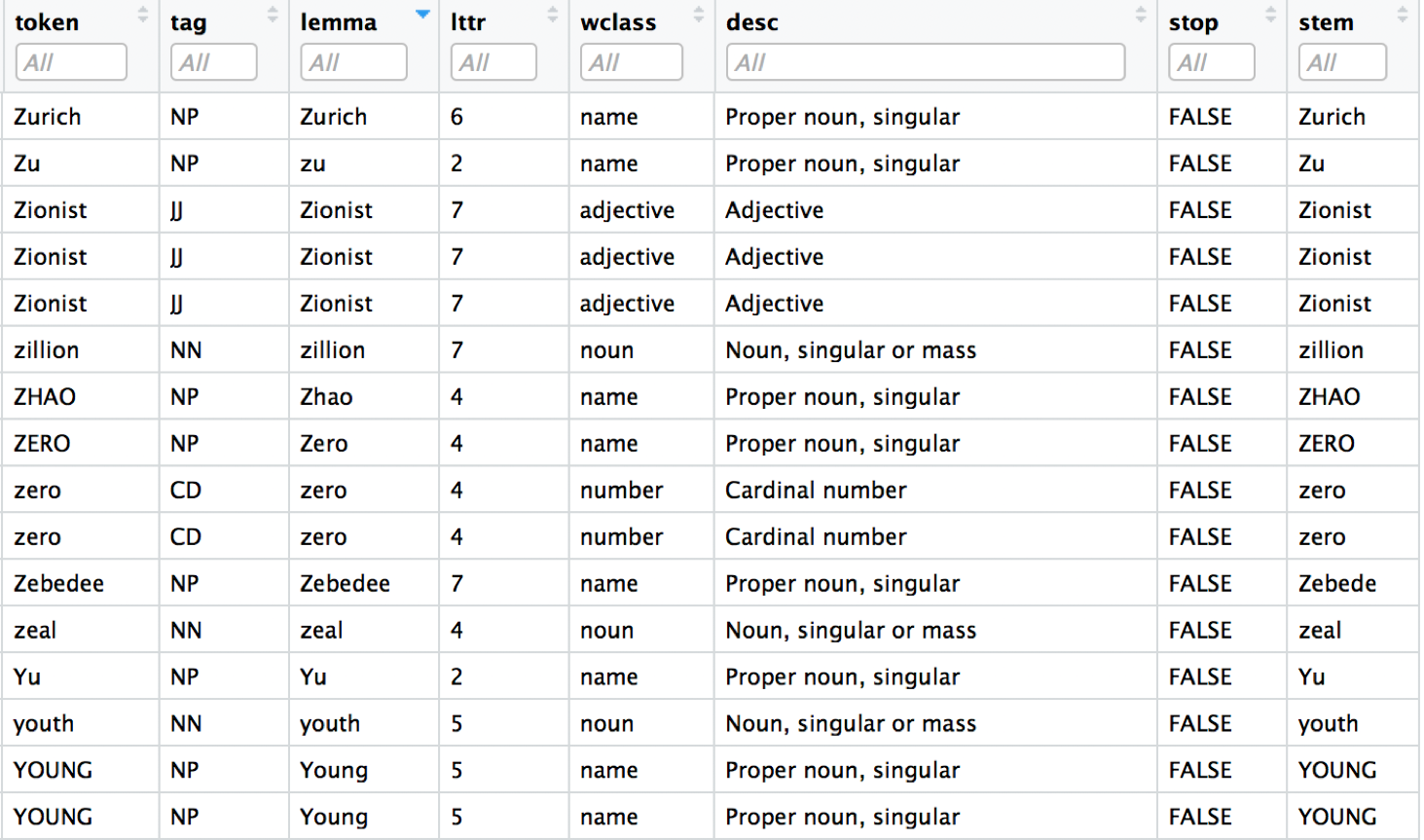
The following tables provide summary information for the training set, including: file layout, missing frequencies, frequencies of affirmative classification (Var=1), total observations and the first 10 observations from the set itself.



Sample training data:

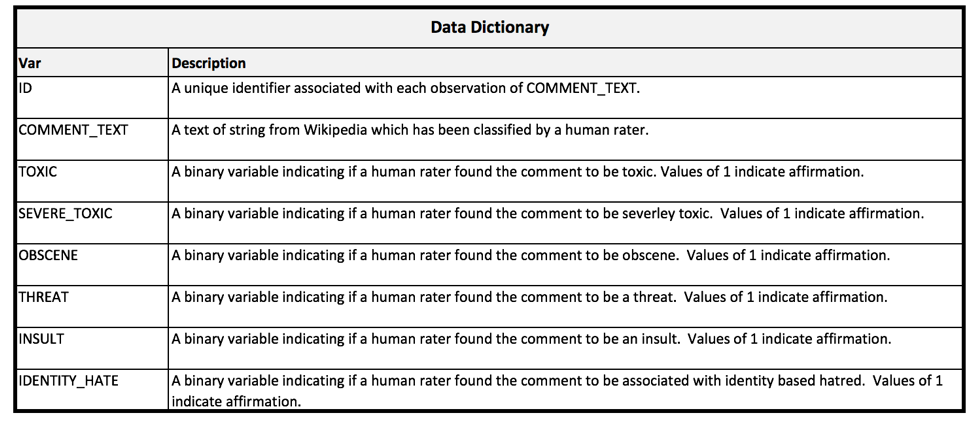


The following table provides a sample of a tagged/tokenized training set for the population of comments classified as affirmative for identity-based hatred.



# Variable List and Data Dictionary

A list of variables in the training set and their metadata is provided in the data dictionary below. The classification is multi-label which means that any particular comment may belong to multiple target classes.



# Data Preparation

Prior to modeling, we must first prepare and analyze the training data. This is done via a Jupyter notebook written in Python so that visualizations can be constructed to quickly display the integrity of the data. Visualizations include:

* **Histograms**

These visualizations display the frequency and distribution for a range of quantitative groups in ‘bins’. Histograms are useful to get a quick overview of the frequency of or ‘shape’ of data in the corpus. In figure 2 we present a histogram of comment text length as well as the natural logarithm of comment text length.

|  |
| --- |
|  |

Figure 2: Histogram of comment text length

* **Box plots**

This visualization displays the distribution and shape of a series of quantitative values for different categories. In this case, we have plotted the distribution of comment length versus a score of toxicity.

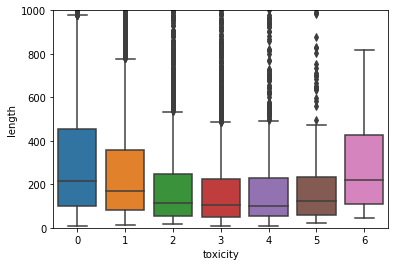
****

Figure 3: Comment text length by toxicity level

* **Bar charts**

This visualization is intended to display quantitative values for different categories. For data preparation purposes, this is useful to determine the term frequency-inverse document frequency relationship of various words in the corpus and to see if they are

|  |  |
| --- | --- |
|  |  |

Figure 4: tf-idf for Obscene, Toxic, Severe Toxic, and Threat labels

* **Word clouds**

A word cloud shows the frequency of individual words used in text data. The higher the frequency the larger the word appears on the word cloud.

|  |  |
| --- | --- |
|  |  |

Figure 5: Word cloud of non-toxic text vs toxic text

* **Heat maps**

This visualization displays quantitative values at the intersection between categorical dimensions. The values in the heat map below represent the correlation between toxicity labels on the x vs the y axes.

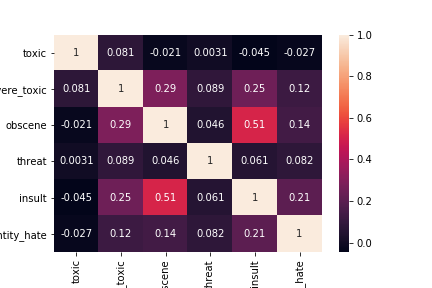
****

Figure 6: Intra-label correlation heat map

With a thorough visual representation of the data, we are able to confirm or reject hypotheses and get a sense for the content of the data. The data preparation step is a necessary step in the value chain of turning raw data into actionable insights.

# Variable Reduction

A text string may be prepared for analysis by representing each word in a vector called a term vector (fig. 2). When the term vectors are tabulated against text records, the resulting term matrix is usually very large and sparse. Removal or consolidation of terms in the term vector may be beneficial for text analysis.



Figure 7: Sample Vector Space Model from (http://blog.christianperone.com/2013/09/machine-learning-cosine-similarity-for-vector-space-models-part-iii/)

Stop word removal is one method of reducing the length of term vectors by removing common terms that do not contribute to the meaning of a phrase. Examples of stop words include “a”, “an”, and “the”. Terms may also be consolidated into equivalence classes: classes of terms which are distinct but have the same conceptual meaning. The terms can be analyzed by transforming the term vector to frequency counts of each term. This frequency term vector can also be reduced by removal of low importance terms as identified by metrics such as the term-frequency-inverse-document-frequency (tf-idf) or by principal components analysis (PCA). Methods such as t-stochastic neighbor embedding (t-SNE) may be used to reduce the vector space to a 2-dimensional space which can be more easily visualized for clustering (fig. 3).

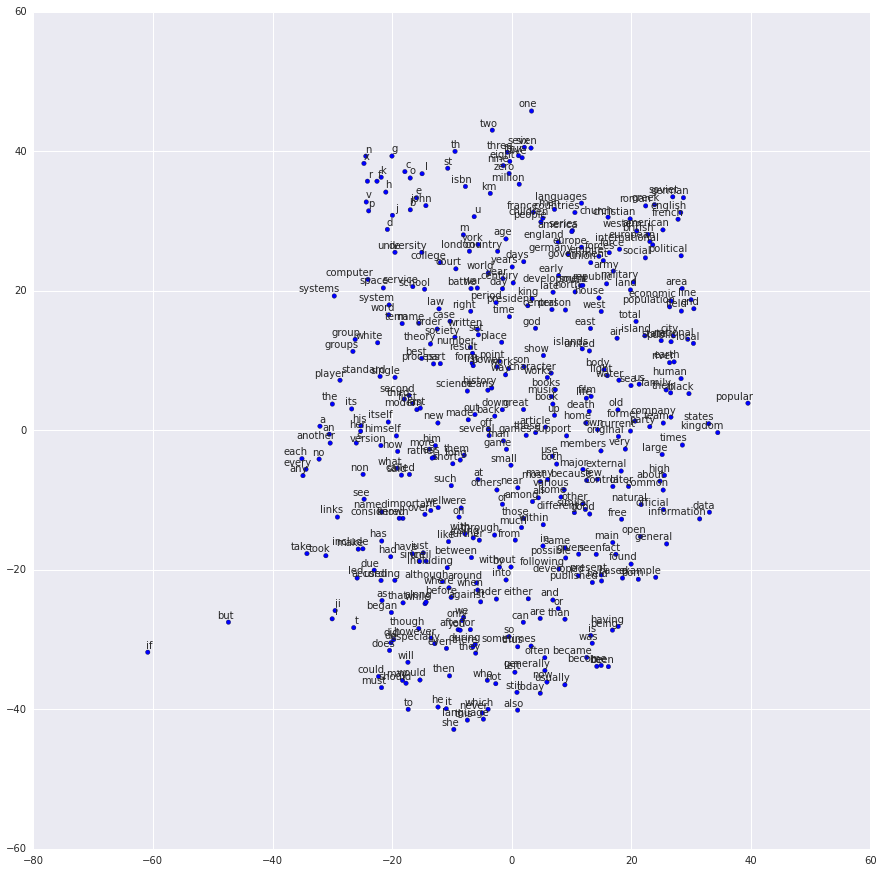


Figure 8: t-SNE Sample Visualization (from https://www.tensorflow.org/tutorials/word2vec)

Alternatively, the Word2vec models may be used to reduce the vector space. These models, developed by Google, are “shallow, two-layer neural networks that are trained to reconstruct linguistic contexts of words”. These models have already been trained on a large corpus and can describe the relative “distance” between terms in a two-dimensional representation (fig. 4).

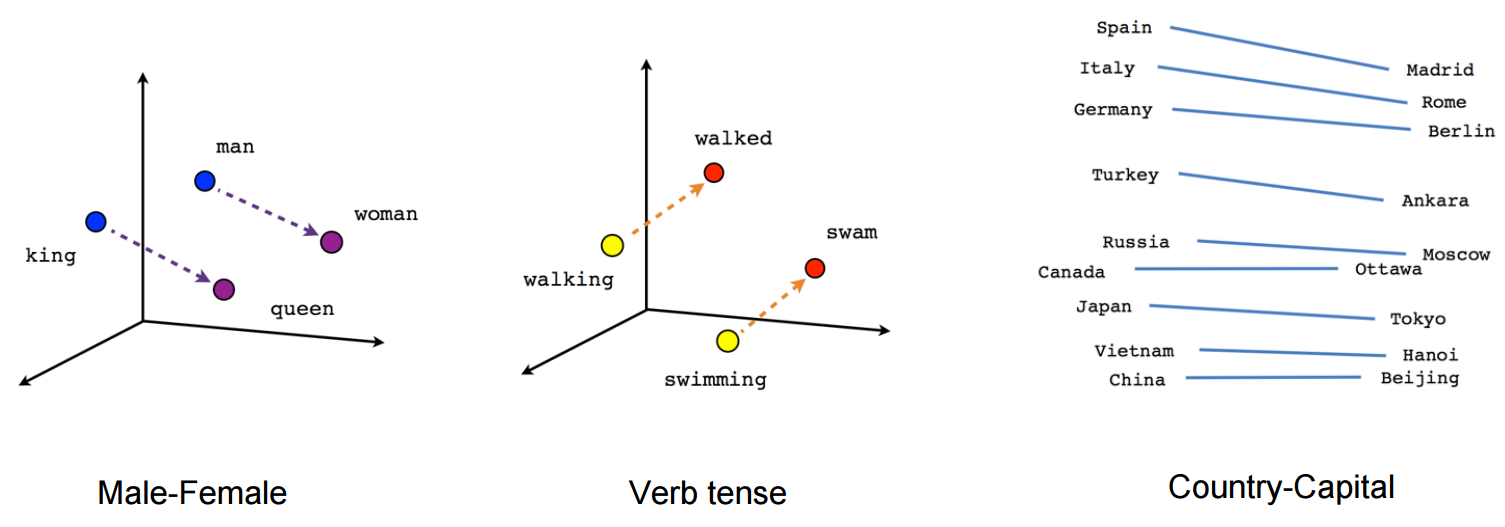


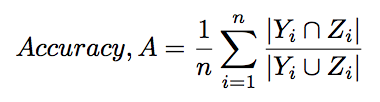
Figure 9: Semantic relationships Sample Visualization (from https://www.tensorflow.org/tutorials/word2vec)

# Measurements & Metrics

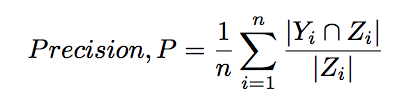
Given that this is a multiclass and multilabel problem there are specific measurements and metrics that apply to model error of this sort. Model selection and evaluation metrics must take into account partially correct, completely incorrect and completely correct sets of labels for each record. In our model evaluation we will explore various metrics including: Exact match ratio, Accuracy, Precision, Recall, F1, and Hamming Loss (Sorower, 2017). For modeling purposes, we will choose our models based on overall Hamming Loss and F1 score, which are explored below. It should be noted that each of these can also be evaluated on individual labels and will be included where relevant when evaluating model error. According to Sorower the definitions for these various metrics are as follows:

**Exact Match Ratio** is simply the ratio of the number of instances whose predicted labels match the actual labels to the total instances.

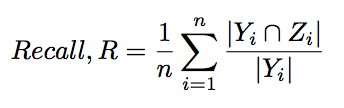
**Accuracy** is the ratio of predicted correct labels to the total predicted plus actual labels averaged across all instances.



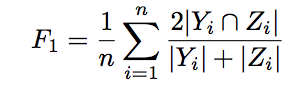
**Precision** is the ratio of predicted correct labels to the total actual labels, averaged across all instances.



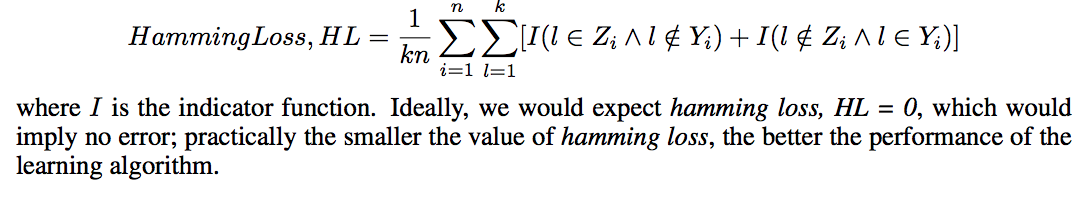
**Recall** is the ratio of the predicted correct labels to the total predicted labels, averaged across all instances.



**F1** is the harmonic mean of Precision and Recall.



**Hamming loss** is based on the number of times a classification is either incorrectly predicted, or missing from predictions made from each model. This forecast of incorrect and missing errors is then normalized over all training examples for an aggregate score:



In order to measure customer engagement, the dashboard should also present customer-centric metrics as defined by the business users. Key metrics that can be measured should include the following as defined by Tawde, 2017:

* **Website Visitors:** The number of unique visitors to the website.
* **Frequency of visits:** How often a user returns to your website.
* **Average Session Duration:** The average amount of time the user spends on the site.
* **Bounce Rate:** the percentage of single-page visits.
* **Visitor Recency:** the amount of time elapsed since the user’s last visit to the website.

# Methodology

Team 4: Green intends to analyze, develop, test and implement an analytical prototype that detects toxic comments in real-time on an aggressive 10-week delivery schedule. This product should be flexible enough to be tailored to various customers’ requirements. The interface will consist of a self-service dashboard used to monitor comments on the customer’s online platforms. This dashboard will give a real-time, aggregated look at all the key leading and lagging indicators of comment toxicity to enable quick mitigating action. It should also provide drill-down capabilities to give the user a more detailed synopsis of toxic comments. This dashboard will implement the predictive analytic classifier trained on the labeled toxic comment dataset provided by the Conversation AI team. This model will also be tested for generalizability on an external dataset.

In order to meet the aggressive schedule and dynamic nature of analytical projects, the team will be organized according to agile principles such as:

• Continuous delivery  
 • Accommodation of changing requirements based on research findings  
 • Close, frequent collaboration of analysts  
 • Focus on analytical best practices

The classification model will be built using Python and code will be controlled and versioned using GitHub. An extensive exploratory data analysis (EDA) will be conducted in order to fully test and refine hypotheses via statistical methods. Any feature engineering and data cleansing required will occur in parallel and iteratively throughout the EDA. A subsequent model selection and tuning phase will test several classification models for performance. After a champion model is selected it will be validated and subsequently implemented into the final dashboard prototype.

# Computational Methods

The Project team will use a number of computational methods to understand the nature of online harassment comments. First and foremost, our data has already been segmented into 7 binary indicators, so each subset can be evaluated in terms of key terms, these can be visualized using word clouds and other plotting methods.

**Data Term Matrix:** Given that our data is a set of comments made by users, we can take a naive approach to try to understand the nature of each binary indicator. For each subset we can take the most mentioned and reoccurring terms and create a matrix, of terms and number of instances. This would yield a vector of word use, by document or comment.

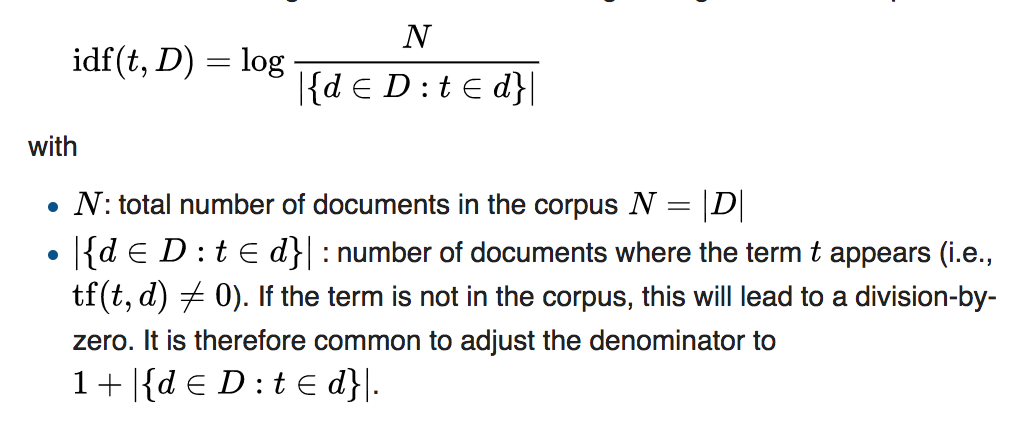
Ideally, we could use a number of text analysis techniques to clean these up and render similar word choices to be equivalent. Stemming, removal of white space and special characters will all be considered in an iterative process in order to create a clean data term matrix for our data set.

**Equivalence Classes/Terms:** Given that multiple terms may be misspelled, worded differently or have special characters additions or subtractions as part of the modeling process, we will explore how certain terms can be grouped together, e.g. ‘Ms.’ versus ‘miss’ should be grouped together and thought of as one term.

**Term Frequency \* Inverse Document Frequency:** Given a large number of terms for our dataset we can then reduce our set of terms to those which are the most unique and significant for classification using a Term frequency \* Inverse Document Frequency weighting system. This will help reduce a large data term matrix down to a reasonable size for analysis. Ideally this would render 20 - 30 top terms for a corpus.

**Term Frequency:** Term frequency or how often a specific word occurs within a comment can be used as a filter to select the most recurring terms across all comments. This is often weighted by some indicator that represents how often a word occurs across the whole corpus. In this case we will use Inverse Document Frequency.

**Inverse Document Frequency:** Inverse Document Frequency is a measure of whether a term is rare or frequent across all comments. IE the word ‘the’ or ‘and’ would have less values versus ‘terrible’ or ‘mine’. It is the log of the number of comments divided by how often the term is within each comment. IE a word that occurs in every document would have an IDF value equal to zero.

[IDF](https://en.wikipedia.org/wiki/Tf%E2%80%93idf#Inverse_document_frequency)

**Euclidean Distance, K Means, and Elbow Plots:** Given a smaller data term matrix, and a large number of comments as observations, we can test our data classification ratings using a k means approach. Ideally an obvious/logical rating system combined with numerical justification would yield a nice set of distinct clusters from which we could model results. IE gender-based harassment distinction versus racially based comments.

In order to test these assumptions given to us within our data, we can use Euclidean distance of our term vectors, elbow plots, and k means to see if our 7 categories are justified by the data.

K-Means classification or k means can be used on our clean data term matrix Euclidean distance values in order to evaluate how related certain comments are to others. Ideally this can be used to choose the number of means or clusters that emerge from the data. Using a k means approach will involve the project team choosing the number of clusters using a heuristic designed to capture a large portion of the variation in the data.

For this project we will use the elbow rule, which looks at the percentage of the variation in the data explained by successive additions of k clusters. IE going from 10 - 11 or 4 to 5 clusters will yield and increase in capture of variation percentage in the data. In our case our cut-off criterion is that if the next successive addition captures significantly less variation in the data. (see below).

**Logistic regression and other Linear Models:** Given that we are satisfied with our selection of term vectors we can then use these vectors as weights within a logistic regression or linear model.

**LSTM and NN approach:** Another approach we will use long term and short-term memory neural nets, various architectures will be considered during the modeling process.

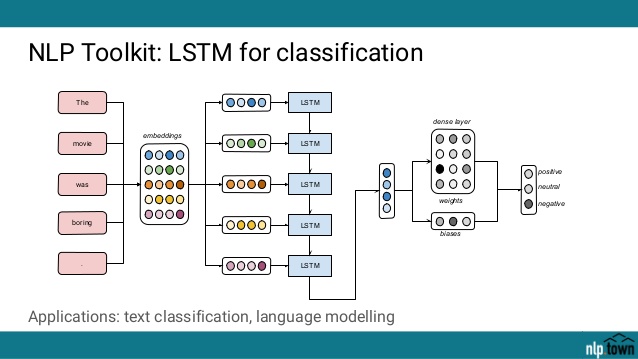


Figure 10: LSTM for classification sample (from https://www.slideshare.net/hendrikdo/yves-peirsman-deep-learning-for-nlp)

# Output Summary

A summary of potential metrics and visualizations for model selection, validation and presentation are provided below.

**Area Under an ROC Curve:** For each class in our dataset we will test how well our model can predict across high and low confidence prediction areas. IE if our model has high confidence, do class predictions occur more often. These can then be plotted in terms of true positive rate and false positive rate. (Fig. 6)

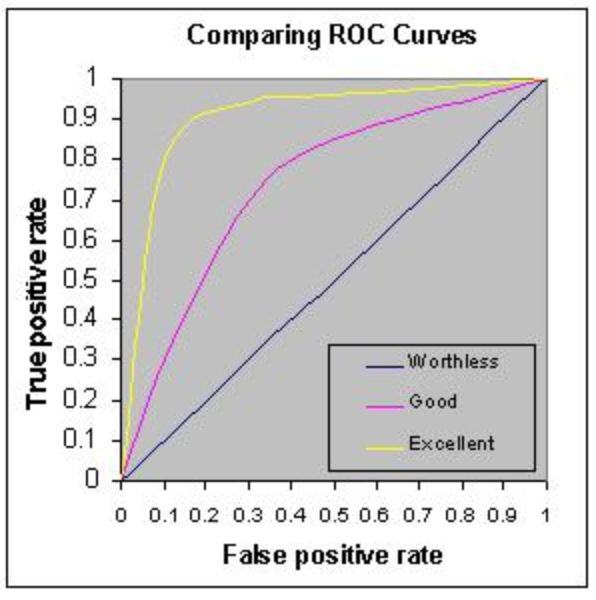


Figure 11: Sample ROC Curve

**Goodness of fit statistics:** including R, R^2, Sample precision and recall, test statistics, and p values will be considered.

**Confusion matrices (k-fold cross validation):** For each class we can evaluate our confusion matrices, these include four boxes: True Positive, True Negative, False Positive and False Negative. For all models we would prefer high True Positive and False Negative Values and low True negative or False Positive values.

**Cluster profiling:** In order to evaluate how prediction variables cluster together we can use distance plots, scree plots and k-means / elbow plots to show overlap of y predictors in terms of our x reference term vectors.

**Elbow Plot:** As we can see below 4-5 k-means clusters would be ideal for this example, as the within groups sum of squares (variation) explained by additional clusters is much smaller, IE looks like an elbow. (Fig. 7).

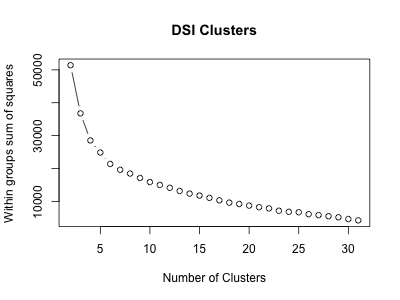


Figure 12: DSI Clusters

# Analytics Project Plan

## Project Management Approach

The project team will follow a CRISP-DM methodology (fig. 8). The design of this document is intended to follow that framework starting with identification of the business problem and defining steps to gain understanding of the business. Subsequently, an iterative step for data understanding and continued interfacing with the business to fully understand requirements and constraints will be conducted. Once satisfied with our initial understanding we seek to explore and understand the data set while preparing it for modeling. This preparation includes the data quality audit report and any subsequent data cleansing and imputation that is required. This exploratory data step also enables our analysts to identify what techniques may be useful during modeling and subsequently implement them for evaluation. During evaluation, the business will once again be engaged in order to ensure that the solution meets all the business requirements. Once the business signs off we will be ready for deployment of the analytical application.

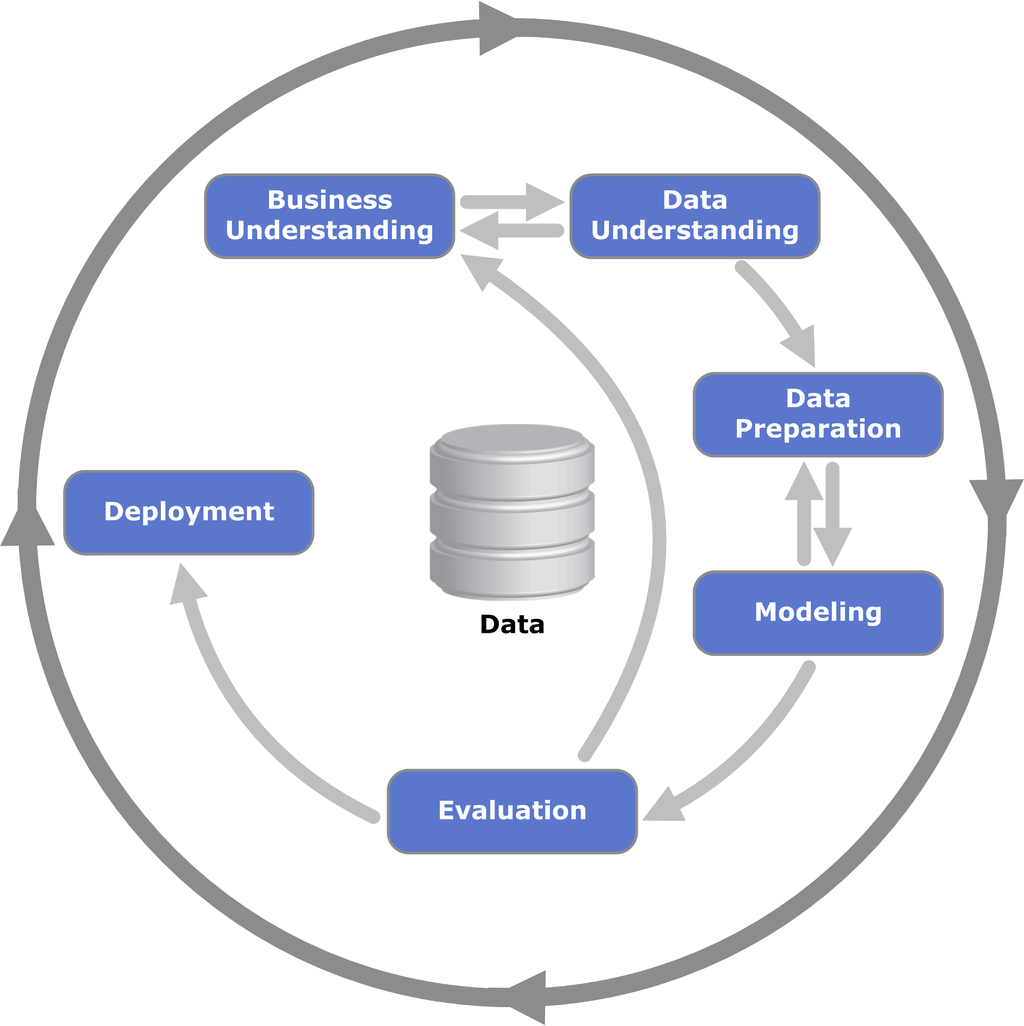


Figure 13: CRISP-DM Methodology

## Project Deliverables

Project deliverables will be due on a biweekly cadence. This will insure adherence to the aggressive 10-week delivery schedule. The deliverables are as follows:

1. *Project Plan and Analysis Plan (due 01/21/2018)*: Seeks to identify the business problem and give an overview of the overall project goals, work plan and team structure
2. *Solution Development and Implementation Report (due 02/04/2018)*: A more technical document that elaborates on the Project Plan and includes variables and sample visualizations
3. *Data Quality Auditing, Cleansing, Organizing and Exploratory Analysis (due 02/18/2018)*: A review of inputs and outputs of the exploratory data analysis as well as technical details of the methods used
4. *Video Presentation (due 03/11/2018)*: A video presentation detailing the analytical model
5. *Final Report (due 03/18/2018)*: Concluding remarks and results from the full project

## Project Technical Requirements

In order for the prototype to provide value to customers and be implementable it must meet the following technical requirements:

1. Must use Python programming language to implement data analysis
2. The hardware and software must pass the IT architect review
3. The dashboard must be accessible 24 x 7 and have a short response time
4. The dashboard must be accessible from any authorized mobile device
5. The platform must adhere to corporate security policy for sensitive data

## Project Constraints

* Acquisition of new data and labeling would delay the project beyond the 10-week deliverable period
* The data types and structure provided to the final dashboards will be limited to those in the Wikipedia comment corpus
* The dashboard will display the following metrics:
  + Toxicity index (defined by the project team)
  + Word cloud of toxic keywords of phrases
  + Toxicity classification and clustering maps
* Interfacing of the dashboard platform with any other platforms is not included

## Project Assumptions

* Scope of the analytic project will not change
* Project plan is approved by the sponsor
* Human labeling of toxic behavior is consistent and universal
* The data pipelines, technology stack, and other resources required to build a complete analytical solution will be available within the budget constraints
* The corpus does not include adversarial examples targeted at bypassing automated moderation
* The data are limited to the English language

## Project Risks

A discussion of this project’s risks is provided below in the form of a series of matrices. The Risk Assessment Matrix identifies the risks, their likelihood, impact and detection difficulty as well as the project phase in which they may exist. The Risk Severity Matrix clusters these risks by likelihood and severity and classifies the overall project risk. The Risk Response Matrix identifies risk mitigation strategies as well as the responsible party for managing the risk.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Risk Assessment Matrix** | | | | | |
| **Risk #** | **Risk Event** | **Likelihood** | **Impact** | **Detection Difficulty** | **When Occurring** |
| R1 | The data provided by distributors is incomplete and/or unreliable | 2 | 5 | 2 | Data Acquisition |
| R2 | Stakeholders are unable to agree on final metrics and other dashboard KPIs | 3 | 3 | 3 | Gathering Requirements |
| R3 | Resources become unavailable due to unforeseen circumstances | 3 | 5 | 5 | Anytime |
| R4 | Technical issues with dashboard | 4 | 4 | 4 | Develop Prototype |

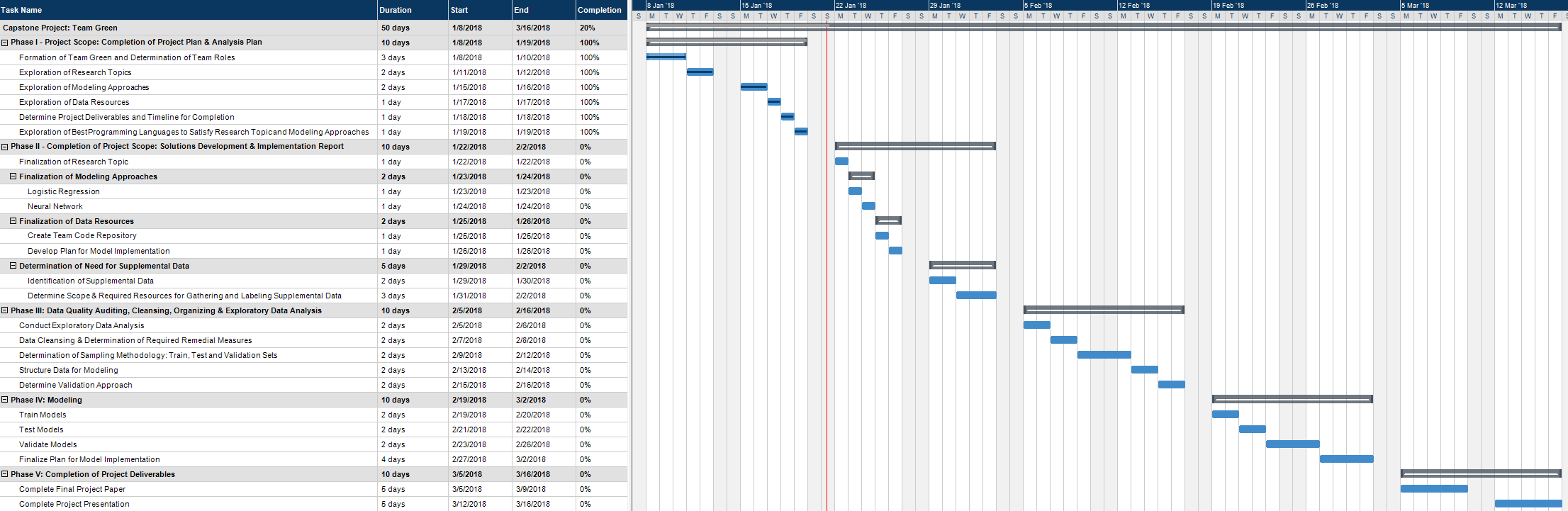
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Risk Severity Matrix** | | | | | | | |
|  | **Likelihood** | | | | | |  |
|  |  | 1 | 2 | 3 | 4 | 5 |  |
|  | 5 |  | R1 | R3 | R4 |  |  |
|  | 4 |  |  |  |  |  |  |
| **Impact** | 3 |  |  | R2 |  |  |  |
|  | 2 |  |  |  |  |  |  |
|  | 1 |  |  |  |  |  |  |
|  |  | 1 | 2 | 3 | 4 | 5 |  |

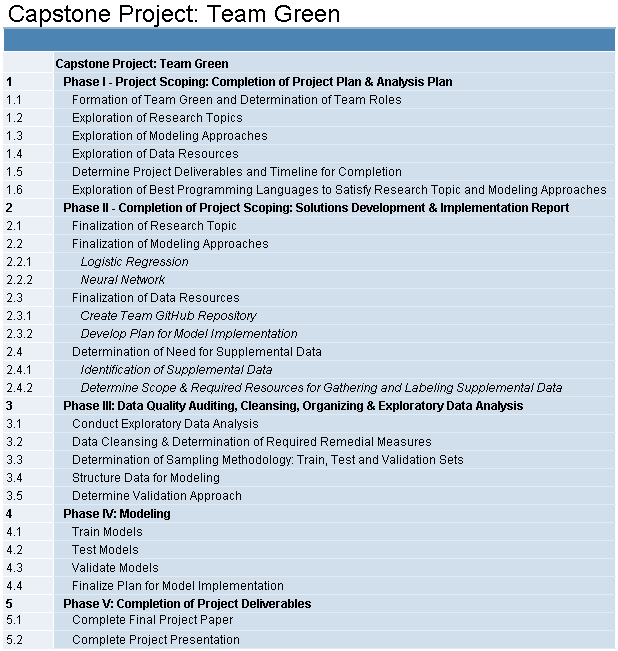
Based on the placement of risks on the Risk Severity Matrix, this project has medium risk. There is one risk event (R4) which is in the high-risk region (red zone). Two risks (R1, R3) also have very high detection difficulty (4+). These risks have medium likelihood but high impact and are also difficult to plan around because they are difficult to detect. The final risk (R2) has medium likelihood and impact and puts the overall risk level for this project at medium.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Risk Response Matrix** | | | | |
| **Risk #** | **Risk Event** | **Response** | **Trigger** | **Responsible Party** |
| R1 | The data provided by distributors is incomplete and/or unreliable | Identify secondary data sources | Exploratory data analysis reveals data issues | David Law |
| R2 | Stakeholders are unable to agree on final metrics and other dashboard KPIs | Collaborate with project sponsor to insure final metrics relate to the business problem | Data QA process shows metrics are unusable | Ricardo Herena |
| R3 | Resources become unavailable due to unforeseen circumstances | Distribute workload evenly | Official personal leave requested | Luis Mesen |
| R4 | Technical issues with dashboard | Remove unnecessary features directly tied to project success | Customer complaint | Joseph Ellis |

## 

## Gantt Chart





## Cost

The Toxic Comment Classification Project is scheduled to begin on 1/8/2018 and to conclude on 3/16/2018. This timeline consists of 50 working days, with an anticipated work day consisting of 4 hours for a total of 200 billable hours. There will be 4 analysts assigned to this project, with an hourly rate of $125.00 per analyst. The anticipated cost for the completion of this project is $100,000, consisting of a combined hourly rate of $500.00.

## Resources

Resource Management will be used to ensure the required human resources are identified and the appropriate roles are assigned. While we anticipate shared responsibilities amongst these different roles, the following resources have been assigned primary responsibility for each of the roles crucial to project success:

**Project Sponsor - Nethra Sambamoorthi**: The main role of the Project Sponsor will be ensuring that all deliverables are met to satisfy the requirements for project completion, while acting as a key resource for any questions and concerns that arise.

**Communications and Engagement Manager - Ricardo Herena**: The Communications and Engagement Manager is primarily responsible for acting as the main point of contact between the project team and project sponsor. Additionally, they are responsible for team engagement and ensuring effective communication amongst all team members. This responsibility includes ensuring that all project deliverables are met on time and to the satisfaction of the project sponsor.

**Programming Languages Analyst - Joseph Ellis**: The Programming Languages Analyst is primarily responsible for ensuring that all programmatic requirements necessary for project completion are met. This includes coordinating all programmatic efforts for the project team through the use of a shared code repository.

**Analytics Modeling Analyst - David Law**: The Analytics Modeling Analyst is primarily responsible for ensuring the appropriate modeling and model validation techniques are utilized by the project team. This responsibility includes supporting the team to ensure successful execution of model development, validation and implementation.

**Report Writing Analyst - Luis Mesen**: The Report Writing Analyst is primarily responsible for ensuring the successful completion of all reporting requirements for the project. This includes coordinating the contributions of the project team as well as ensuring effective written communication so that all project deliverables are clearly articulated.

# References

Bakharia, A. (2017, November 27). Using TSNE to Plot a Subset of Similar Words from Word2Vec. Retrieved February 02, 2018, from <https://medium.com/@aneesha/using-tsne-to-plot-a-subset-of-similar-words-from-word2vec-bb8eeaea6229>

Duggan, M. (2017, July 11). 1. Experiencing online harassment. Retrieved January 21, 2018, from <http://www.pewinternet.org/2017/07/11/experiencing-online-harassment/>

Our Latest Update on Safety. (n.d.). Retrieved January 21, 2018, from <https://blog.twitter.com/en_us/topics/product/2017/our-latest-update-on-safety.html>

Rep. Katherine Clark tackles online crime aimed at women and girls. (n.d.). Retrieved January 21, 2018, from <https://thinkprogress.org/rep-katherine-clark-tackles-online-crime-aimed-at-women-and-girls-bfbdcb8805b1/>

Sorower, M. S. (2010). A literature survey on algorithms for multi-label learning. Oregon State University, Corvallis, 18.

Tawde, S. (2016, May 30). Web Analytics engagement, Metrics, user engagement, web analytics, website traffic. 7 Important User Engagement Metrics For Your Website. Retrieved February 02, 2018, from <http://www.digitalvidya.com/blog/7-important-user-engagement-metrics-for-your-website/>

TensorBoard: Embedding Visualization | TensorFlow. (n.d.). Retrieved February 02, 2018, from <http://www.tensorflow.org/versions/r0.12/how_tos/embedding_viz/>

Young, S. (2017, April 04). Online Harassment, Digital Abuse. Retrieved January 21, 2018, from <https://datasociety.net/blog/2017/01/18/online-harassment-digital-abuse/>