Data Impact Challenge II Answer Submission Template

- **Question:** Youth Mental Health
- **Team name and list of all team member names**:
  SAS – Jos Polfliet, Marie Soehl, Greg Horne, Tim Trussell, Antoni Dzieciolowski

Data and Analysis

- **Data Custodian Organizations and Data Sources**
  - ✓ Twitter (https://twitter.com/)
  - ✓ PAN (http://pan.webis.de/index.html)
- **List of Datasets Used**
  - ✓ Twitter: Data accessed through Twitter API (https://dev.twitter.com/rest/public)
  - ✓ PAN: Author Profiling Challenge data (http://pan.webis.de/clef16/pan16-web/author-profiling.html)
- **Exclusions**
  - ✓ Tweets without a public location were omitted.
  - ✓ Tweets outside of Canada were omitted.
  - ✓ Tweets having a low likelihood of being within the 13-17 age ranged were omitted.
- **Nature and Size of Cohort**
  - ✓ Original sample pool of tweets from all ages published within Canada between January 6, 2016 and January:
    - 285,595 tweets
    - 24,605 users
  - ✓ Full pool of all tweets from 24 605 users:
    - 2,283,232 tweets
  - ✓ Final analysis pool of Canadian users age 13 – 17:
    - 1.1 million tweets
    - 926 users
- **Data Timeframe**
  - ✓ Tweets originally pulled, published in Canada, were from January 6, 2016 – January 13, 2016. All tweets from the users in the original pool were used in the analysis, regardless of date.

Summary of the Analysis Methodology

- The twitter API was used to collect a sample of tweets published within Canada between January 6, 2017 and January 13, 2016.
- All tweets from the users in the aforementioned sample were pulled using the API.

- PAN data was used to build a predictive model for age. The word vectors were converted to numerical representations by using bag-of-word conversion and singular value decomposition next. Neural networks, random forests, decision tree and logistic regression were fitted to the data. Random forest had the best performance and was used to do the final prediction. The model was built in SAS Enterprise Miner and used default settings for the built-in Text Miner and Random Forest nodes. The model was used on the twitter dataset to identify tweets published by the 13 – 17 age demographic.
- Text analysis was run on the tweets by the 13 – 17 year olds. Tweets were classified according to a manually defined taxonomy containing keywords related to the topics of interest. Conjugations, stemming of words and synonyms were added automatically by the SAS Contextual Analysis Software.
    - ✓ Text topics: feelings, bullying, suicide
- The results were visualized and topics were analyzed across all tweets and users in the youth sample pool.

Description of Findings

- **Numerator and Denominator**
    - ✓ Numerator: number of tweets talking about suicide and bullying by Canadian youth aged 13 - 17
    - ✓ Denominator: All tweets by Canadian youth aged 13 - 17
- **Key Statistics**
    - ✓ Natural language processing: bag-of-words, singular value decomposition
    - ✓ Predictive modeling: neural networks, random forests, decision trees, logistic regression
    - ✓ Text mining: classification method
    - ✓ Data visualization
- **Summary of the Findings**
    - ✓ It was found that 1% of tweets were about bullying and 0.035% of tweets were about suicide.
- **Key Limitations**
    - ✓ Twitter does not provide the age of the user. Other social media sites or blogs, including Facebook, require a birth date when signing up. Using these sources may have other drawback, but eliminate the need for an age prediction model.
    - ✓ The best submission for the PAN author age prediction had a misclassification rate of 40.53 %, while our model had a misclassification rate of 41.12 % (on the validation data set). There is little room for improvement using the PAN data.
    - ✓ The age prediction model did not include emoji characters. Including these could potentially enhance the model.
    - ✓ Tweets published in Canada may not be from Canadians.
    - ✓ The Twitter API has restricted collection rates. It took 3 days to download 2.3M tweets.

- ✓ The analysis did not explore sentiment of tweets. This type of analysis could strengthen the findings. This could be performed in the same tool used for text mining.
- ✓ The classification of topics faced limitations due to teen jargon, such as, “This is sooo funny, I am killing myself laughing” being considered a tweet about suicide. Further manual tweaking of the classification model would be required for improvements.

- **Additional Insight**
  - ✓ 10% of tweets were about feelings. There was a 39% correlation between suicide and bullying. 12.7% of users in the youth sample pool have tweeted at least twice about suicide.
- **Implications of Analysis**
  - ✓ The analysis provides insight into bullying and suicide tweets of a relevant demographic over time.
  - ✓ Analysis can be done without survey costs and consent requirements.
  - ✓ It allows the ability to monitor the changing ratio of tweets within a specific topic in response to program or marketing initiatives.
  - ✓ The analysis can be used to identify high risk teens or communities and appropriate action can be taken.
  - ✓ If the analysis can be associated with suicide occurrences, predictive modeling can be used in conjunction with the text analysis to determine specific factors around the timeline and likelihood of individuals to harm themselves.