**CURO Poster Guide**

**Quick Pitch:** The word vaccination has caused controversy since its inception, and these controversies still continue today. Despite the fact that vaccination provides the most effective method of preventing infectious diseases, opponents still question the effectiveness, safety, and necessity of recommended vaccines. While the effectiveness and safety of vaccines has been widely studied and verified, the anti-vaccine movement has still brought forth vaccine hesitancy, or a delay in acceptance or a refusal of vaccine services. It is an ever-growing and constantly changing problem that needs constant surveillance. The Internet plays a large role in disseminating vaccine misinformation to a large number of people, which contributes to the vaccine hesitancy problem. In order to combat this spread of misinformation online, it is important to first recognize true facts from false ones. We attempt to develop a machine learning strategy using natural language processing (NLP) that allows one to identify misinformation in vaccine-related webpages. We aim to meet this goal through the use of the low-dimensional document embedding algorithm, Doc2Vec, as well as semi-supervised classification. We take a small sample of manually labeled vaccine webpages and a large amount of unlabeled vaccine webpages, and attempt to classify misinformed webpages from accurate ones using pairwise cosine similarities of document vectors. We find that Doc2Vec serves as a reasonably good starting point for document classification and, ultimately, a semi-supervised framework for identifying false or misleading documents. The more important question remains of what truly makes a document reliable or not. We gain a few insights from some of our visualizations in Figures 2 & 3, but any future works should look to expand upon this issue.

**NLP:** Natural Language Processing refers to a field in Computer Science and computational linguistics concerned with the programming of computers to effectively process and understand natural language text. Much of the NLP work being done today has relied heavily on machine learning, in which programmers rely on using statistical inference to automatically learn “rules” of the language corpora.

**Semi-Supervised Classification:** Semi-Supervised Classification uses both labeled and unlabeled data in order to predict the classification of unseen data. For our project, the labeled data consists of 20 manually labeled vaccine webpage documents as either TRUE or MISINFORMED. The unlabeled data consists of 1095 vaccine webpage documents collected through the use of Google’s Custom Search API. We then attempt to infer the label of the unlabeled examples in advance before building the classifier (transductive learning). This is accomplished through the use of the Doc2Vec algorithm.

**Doc2Vec:** Doc2Vec is a low-dimensional document embedding algorithm that represents a document in vector space. Doc2Vec was created by a team of researchers led by Tomas Mikolov at Google. While Doc2Vec is not the only approach for document embedding, there have been many claims that it outperforms other embedding schemes, such as Bag of Words or LDA (<https://arxiv.org/pdf/1507.07998.pdf>). Doc2Vec representations are trained to be useful for predicting words in a paragraph. More precisely, it concatenates the paragraph vector with several word vectors from a paragraph and predicts the following word in the given context. Both word vectors and paragraph vectors are trained by the stochastic gradient descent and backpropagation. While paragraph vectors are unique among paragraphs, the word vectors are shared. At prediction time, the paragraph vectors are inferred by fixing the word vectors and training the new paragraph vector until convergence.

**Cosine Similarity:** A measure of similarity between two non-zero vectors of an inner product space by measuring the cosine angle between them.  It is a measure of orientation and not magnitude: two vectors with the same orientation have a cosine similarity of 1, two vectors at 90° have a similarity of 0, and two vectors diametrically opposed have a similarity of -1, independent of their magnitude. It also one to see how semantically similar two document embeddings are.

**tSNE:** t-distributed stochastic neighbor embedding (t-SNE) is a machine learning algorithm for dimensionality reduction that is well-suited for embedding high dimensional data into two dimensional space, which can they be visualized in a scatter plot. It attempts to model each point in such a way that similar objects are modeled by nearby points and dissimilar objects are modeled by distant points.

**Logistic Regression:** Logistic regression measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a logistic function. It makes its predictions based on conditional probabilities of Class | Feature Vector.

**Naïve Bayes:** Naïve Bayes is a simple probabilistic classifier that applies Bayes Theorem with strong independence assumptions between the features. It is a popular baseline for text categorization.

**SVM:** SVM models the problem by attempting to project the data into a higher dimensional space in which the data is linearly separable. SVM then searches for the optimal separating hyperplane, by finding the hyperplane as far as possible from data points from each category.

**Random Forest:** Random Forests is a method that can be used for classification that operates by creating a multitude of decision trees from the input and outputting the class that is the mode of the classes. This usually corrects for decision trees tendency to over fit the data. Decision trees are simply predictive models that map observations about an item to conclusions about item’s target value.

**K-Nearest Neighbors:** kNN is among the simplest machine learning algorithms. The input to kNN considers the k closest training examples in the feature space, and the object is then classified by a majority vote of its neighbors, with the object being assigned to the class most common among its *k* nearest neighbors.