INTRODUCTION: SENTIMENT ANALYSIS

- Extracting attitudes within text
- Converting unstructured information into structured data
- Part of the larger Natural Language Processing field

  - Attitudes: liking, loving, hating, valuing, desiring
  - Emotion: angry, sad, joyful, fearful, admired, proud
  - Mood: cheerful, gloomy, irritable, listless, depressed, buoyant
  - Interpersonal Stance: friendly, cold, warm, supportive
  - Personality Traits: nervous, anxious, reckless, moody, hostile, envious, jealous

INTRODUCTION: TASKS RELATED TO SENTIMENT ANALYSIS

- Classification
  - Positive, Negative, Neutral
- Strength
  - Score on a scale of -1 to 1
- Detect the target of the sentiment
- Classify into more complex affective states

One problem here is that the company has only two enrollment dates for the 401K plan, January 1 & July 1. In my case, I started working in January so I wasn’t eligible for enrollment until July of the following year. I had to wait 18 months before they started contributing. There is no valid reason to only have two enrollment dates. An employee should be enrolled as soon as they have completed their eligibility requirements. However, a 6% employer contribution is pretty good, but it would be a hell of a lot better if was on total pay and not just base pay.

Example Text: About 401K plan

Data and Scalability

- People don’t like voting. They like expressing views in their own words
- Allows for real-time, granular analysis
- Creation of metrics and comparisons
WORKERSHIP INTRODUCTION TO SENTIMENT ANALYSIS

INTRODUCTION: APPROACHES

Rule Based Approach

- Lexicon of Positive and Negative Words
- Manually Crafted Rules based on linguistic expertise

Machine Learning Approach

- Supervised Learning
  - Comment that needs to be classified
  - Build a prediction model to classify any text into sentiment classes
  - Rule based techniques useful for feature generation

Machine Learning Approach

- Supervised Learning
  - Comment that needs to be classified
  - Supervised Learning
  - Make a predictions based on the training model

INTRODUCTION: APPROACHES

Machine Learning Approach

- Supervised Learning

Build a prediction model to classify any text into sentiment classes

Process: Sentiment Analysis Process

- Data Cleaning and Tokenization
- Feature Extraction
- Training
- Testing and Accuracy

Process: Data Cleaning and Tokenization

- Examples of data cleaning
- HTML Text Cleaning
- Capitalization
- Repeated Characters
- Punctuations, whitespace, stop words
- Platform specific symbols
- Negation
- Regular Expressions are very helpful to search by patterns
- Break apart the text into tokens

Made by

1

2
WORKERSHIP: INTRODUCTION TO SENTIMENT ANALYSIS

PROCESS: FEATURE EXTRACTION

- Rule Based Features
- Bag of Words, N-grams
- Dimension Reduction Techniques
- Word Embeddings and Word2Vec

### Bag of Words Approach

- Create a document term matrix
- Columns are all the unique terms in all the comments
- Each comment becomes a vector of term counts
- Using TF-IDF (term frequency-inverse document frequency) to weight importance of a word in the collection of documents
- The columns (all unique terms) act as features

<table>
<thead>
<tr>
<th>Column</th>
<th>Dog</th>
<th>Love</th>
<th>Hate</th>
<th>Sentiment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dog</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>Positive</td>
</tr>
<tr>
<td>Love</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>Positive</td>
</tr>
<tr>
<td>Hate</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>Negative</td>
</tr>
</tbody>
</table>

### N-grams

- Contiguous sequence of n-tokens
- Bigrams: I like, like dogs, dogs and...
- Trigrams: I like, like dogs, like dogs and...
- N-grams take more word context into account

### Dimension Reduction Techniques

- Singular Value Decomposition can help get rid of redundant data by reducing the dimension of the document-feature matrix

<table>
<thead>
<tr>
<th>Document-Concept</th>
<th>Concept Strength</th>
<th>Concept-Term</th>
<th>Document-Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>U</td>
<td>D</td>
<td>V</td>
</tr>
</tbody>
</table>

Latent Semantic Analysis

Words used in similar contexts have similar meanings

### Word2Vec

- Uses a 2 layer neural network to create word vectors
- Set the hidden layer size to the dimension of the word vectors
- Train using all of the text in the corpus to predict surrounding words
WORKERSHIP INTRODUCTION TO SENTIMENT ANALYSIS

**PROCESS: FEATURE EXTRACTION**

- **Word2Vec**
  - Create a word vector space using the weights for the hidden layer
  - Compare similarity between words using the word vectors
  - Use word embedding vector space as features when you train

**PROCESS: TRAINING**

- Build a prediction model based on the documents and features
  - Logistics Regression
  - Naive Bayes
  - Decision Trees
  - Support Vector Machine
  - Neural Network
- Decisions about which training model to use depends on number of examples, features, computational complexity

**PROCESS: TESTING**

- Apply the model generated by the training process to predict sentiment for any text
- To compute accuracy, divide tagged dataset into training and test sets
- Testing Metrics:
  - **Accuracy**: \( \frac{(\text{True Positive} + \text{True Negative})}{\text{Total}} \)
  - **Precision**: \( \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \)
  - **Recall**: \( \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \)

**BIG PICTURE: SENTIMENT ANALYSIS AND NATURAL LANGUAGE PROCESSING**

- Converting unstructured data to structured data
- Language Ambiguity
- Combining linguistics with computational statistics

Workership
Introduction to Sentiment Analysis

CS 410, UMass Boston
April 11, 2019