Multilateration-Powered Social Media Classification

Brian Morales

April 24, 2023

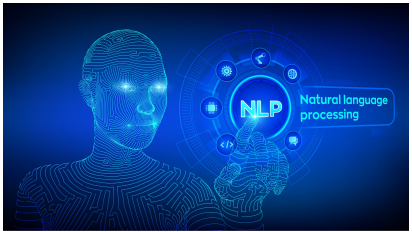
Department of Applied Mathematics

Applied Mathematics
UNIVERSITY OF COLORADO **BOULDER**

Outline

- Natural Language Processing (NLP)
- The Jaccard Distance
- What are Resolving Sets?
- The Information Context Heuristic (ICH)
- Probabilistic Approach to Find Resolving Sets
- Embedding Sentences
- SVC Applications

Motivation



- **Natural Language Processing (NLP)** is the ability of a computer to "understand" human language as it is spoken and written.
- NLP is important for sentiment analysis and bot detection, and we would like to apply NLP's to distinguish twitter post.
- Let X be a **set of words** (e.g. all English words), and 2^X, the **power-set** of X.
- In this work, we represent English sentences as **bag of words** i.e., elements of 2^X.

The Jaccard Distance

- Let $A, B \in 2^X$.
- The **Jaccard distance** between A and B is defined as

$$d(A,B) := 1 - \frac{|A \cap B|}{|A \cup B|}.$$

Example

If $A = \{god, queen, royal\}$ and $B = \{queen, royal, castle\}$ then

$$d(A,B) = 1 - \frac{|\{queen, royal\}|}{|\{god, queen, royal, castle\}|} = 1 - \frac{2}{4} = \frac{1}{2}$$

What are Resolving Sets?

Definition: Resolving Set

$R = \{r_1, \ldots, r_k\} \subset 2^X$ (i.e., a collection of subsets of X) **resolves** the metric space $(2^X, d)$ **if every** $A \in 2^X$ **can be uniquely represented by the** k**-dimensional vector of distances:**

$$\big(d(A, r_1), \ldots, d(A, r_k)\big).$$

In a sense, the elements in R act as "landmarks" to identify any element in 2^X by its distance to those landmarks.

Resolving Set Example

Jaccard distance	$\{1,3\}$	$\{2,3\}$
$\emptyset$	1	1
$\{1\}$	$1/2$	1
$\{2\}$	1	$1/2$
$\{3\}$	$1/2$	$1/2$
$\{1,2\}$	$2/3$	$2/3$
$\{1,3\}$	0	$2/3$
$\{2,3\}$	$2/3$	0
$\{1,2,3\}$	$1/3$	$1/3$

If $X = \{1,2,3\}$ then $R = \big\{\{1,3\},\{2,3\}\big\}$ resolves 2^X because no two rows are identical over the last two columns.

The Information Context Heuristic (ICH)

- The **ICH** algorithm can find nearly optimal resolving sets in a finite metric space utilizing a distance metric.
- The **ICH** becomes unrealistic when X is large because it requires the distance matrix (dimensions: $2^{|X|} \times 2^{|X|}$) to be stored in memory.
- Therefore, the **ICH** becomes rapidly infeasible as the number of words increases.
- **How to find resolving sets of $(2^X, d)$ when X is large?**

Probabilistic Approach to Find Resolving Sets

- Construct a random set $R = \{r_1, \ldots, r_k\}$, where each word is added to r_i with probability 1/2.

Theorem

For all X large enough and constant $C > 0$, $R = \{r_1, \ldots, r_k\}$, with

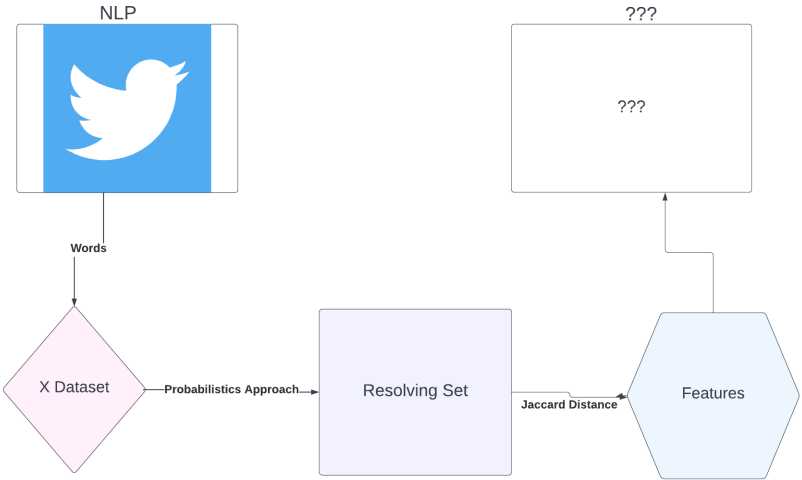
$$k \sim \frac{C \cdot |X|}{\ln |X|},$$

resolves all bags of words of equal size with overwhelmingly high probability.

Embedding Sentences

- Let X be the set of words in a dataset of sentences.
- Each word has one of three sentiments (tones): **positive**, **negative**, and **neutral**.
- We assembled random synthetic sentences with varying proportions of tones.
- Each sentence is represented as an element of 2^X, which we resolved using a set R of cardinality $|R| = 325$, i.e. there are 325 "landmarks."

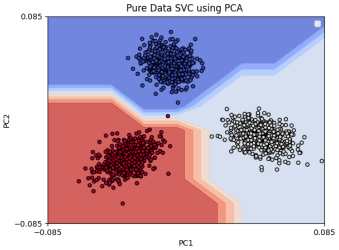
Pipeline Diagram



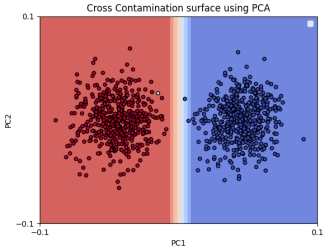
Support Vector Classifier (SVC)

- SVC is a classification model that uses hyperplanes to separate data-points into different classes.
- We applied an SVC to our synthetic sentences after representing each sentence by its 325 features, obtained from the resolving set.
- We used PCA to visualize the performance of the SVC model. (The accuracy difference between the full model and reduced model was minimal.)

SVC - PCA 2D Visualization



Pure Data[1]

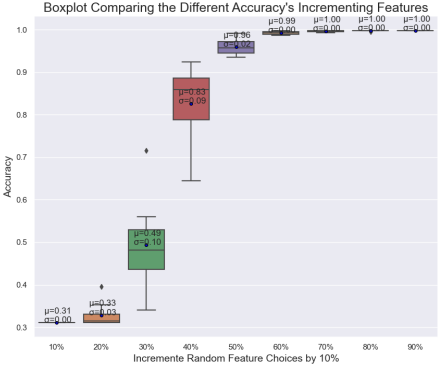


Cross-contamination[2]

[1]SVC plot of the synthetic sentences projected onto PC1 and PC2. Accuracy was rated at 99.07%.

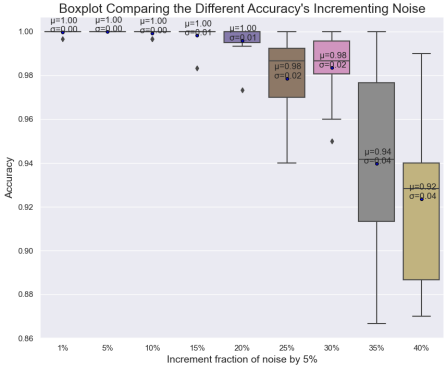
[2]SVC plot of dual tone with 10% cross-contamination, projected onto PC1 and PC2. Accuracy was rated at 99.06%.

SVC - Incrementing Features



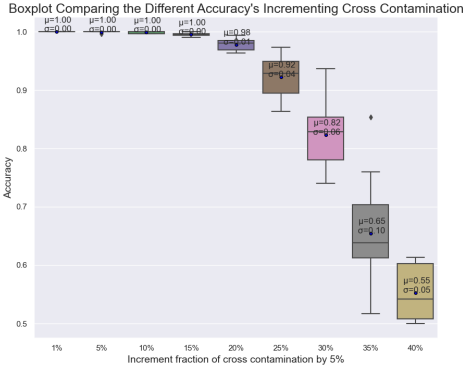
- The total number of words we have to choose from is $n = 1050$, and each sentence length is
  $$S \sim Poisson(\lambda = 50).$$

SVC - Addition of Noise



- Increment the amount of noise of each sentence.
- Noise refers to the amount of neutral words in a sentence.
- Similarly, the number of words being used is $n = 1050$, and $S \sim Poisson(\lambda = 50)$.

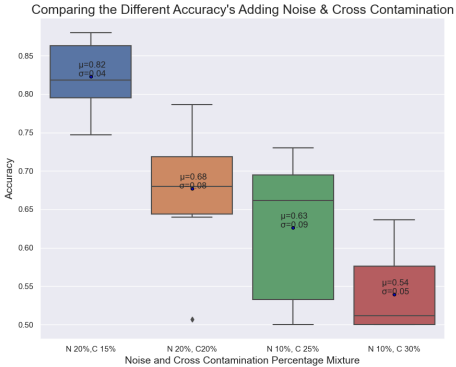
SVC - Cross-contamination



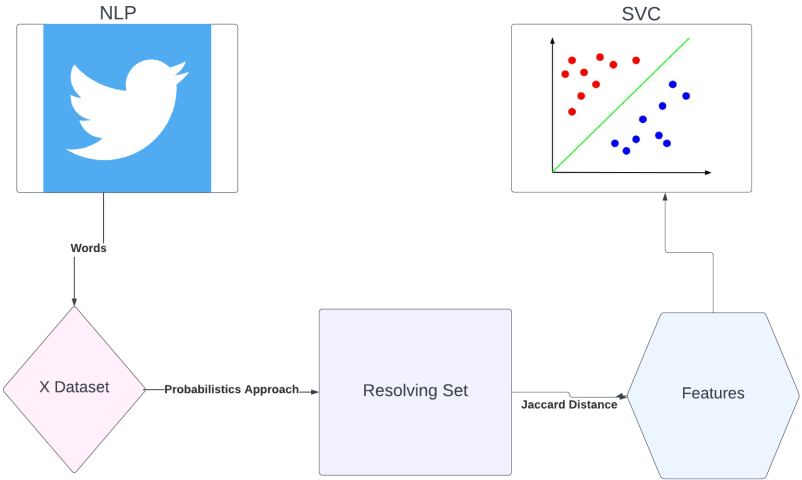
- Increment the amount of negative words of each positive sentence.
- $n = 1050$.
- $S \sim Poisson(\lambda = 50)$.

SVC - Cross-contamination & Noise

- We add a mix of both cross-contamination and noise to make the sentences more realistic.
- $n = 1050$.
- $S \sim Poisson(\lambda = 50)$.



Pipeline Diagram



Future Work

- Improve the embedding's robustness to cross-contamination and noise, as well as sentence length.
- Apply to real-world tweets with application in both sentiment analysis, threats, and bot detection.
- Utilize SVC's to classify new, unobserved tweets.

Acknowledgments

- Manuel Lladser (Research Advisor)
- Alexander J. Paradise (Co-mentor/Collaborator)
- Lladser Research Group

Thank You

Questions?

Backup slides - Jaccard Matrix

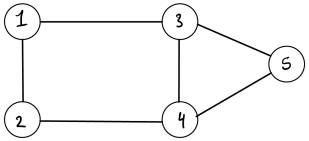
- Example.

Jaccard distance	$\{1,2\}$	$\{2,3\}$	$\{1,2,4\}$
$\emptyset$	1	1	1
$\{1\}$	1/2	1	2/3
$\{2\}$	1/2	1/2	2/3
$\{3\}$	1	1/2	1
$\{4\}$	1	1	2/3
$\{1,2\}$	0	2/3	1/3
$\{1,3\}$	2/3	2/3	3/4
$\{1,4\}$	2/3	1	1/3
$\{2,3\}$	2/3	0	3/4
$\{2,4\}$	2/3	2/3	1/3
$\{3,4\}$	1	2/3	3/4
$\{1,2,3\}$	1/3	1/3	1/2
$\{1,2,4\}$	1/3	3/4	0
$\{1,3,4\}$	3/4	3/4	1/2
$\{2,3,4\}$	3/4	1/3	1/2
$\{1,2,3,4\}$	1/2	1/2	1/4

If $X = \{1,2,3,4\}$ then $R = \big\{\{1,2\},\{2,3\},\{1,2,4\}\big\}$ resolves 2^X because no two rows are identical over the last three columns. R was found using the **Information Content**

Backup slides - Resolving Set Example

X = {1, 2, 3, 4, 5 }



R = { 4, 5 }, (1,1) = 3, (1,2) = 2, (2,2) = 1,
(1, 0) = 5, (0, 1) = 4