Sarcasm Detection on short text
Sai Prashanth Ravichandran | sr6229@rit.edu
Advisor: Dr. Carol Romanowski
Rochester Institute of Technology

Introduction
With the advent of social media, opinion mining on the unstructured text data provides valuable insights.

Motivation: Sarcasm detection is a challenging and interesting problem in text analysis. Since a sarcastic statement can contain positive words with latent negative meaning, it causes issues in sentiment analysis.

Limitations: No context, Sparse features and High dimensionality

Objectives: Understand and implement a system for sarcasm detection using different text modeling techniques

Data
- Hashtag based data collection: #sarcastic, #not, #sarcasm
- Twitter Archiver[1] used to collect data and added to corpus.
- The data set contains around 120k regular and 25k sarcastic clean tweets.

![Fig 1: A list of top 5 Emojis in sarcastic and regular tweets](image)

![Fig 2: A word cloud of top 150 words in sarcastic and regular tweets](image)

Methodology

Data Cleaning
- Removed Retweets, tweets that contain hyperlink, images and in other languages.
- Removed numbers, abbreviations, hashtags and stopwords and converted to lowercase.
- Tokenized, converted contractions and corrected common spelling errors.

Feature Generation
- Unigrams and Bigrams
- Emoticons and Emojis
- Pos-Tagging
- Sentiment
- Punctuations and Capitalization
- Applied TF-IDF Transformer

Classification Algorithms and Metrics
- ML Algos: Multinomial Naïve Bayes, Logistic Regression, Linear SVC
- Metrics: Accuracy - Fraction of correctly classified instances

Data Sample
- Been the most happening trip to mumbai so far. #sarcasm
- So I got in a car accident today... that was fun #sarcasm
- I love spending Saturday nights doing nothing!! Hell yeah!! #SoMuchFun #sarcasm
- 13 hour shift tomorrow... let's go! #sarcasm
- This is the best day of my life! 😊😊😊
- I am liking the direction that everything is going lately 😊😊😊😊

Fig 3: A flow diagram that describes the methodology

Results

Fig 4: Comparison of the three algorithms used

Challenges
- Class imbalance - Used different metrics to counter this.
- Had to deal with different kinds of text encoding.
- Scraped websites to find spelling errors, Emojis and emoticons.

Conclusion and Future work
- The Experiments display that SVM is the most optimum classifier with an accuracy of 89.86% on test data.
- Cleaning played a very important role in improving the model.
- This model can be improved by taking context into account.
- Spell checker can be used.

References