# Multi-class Claims Activity Classification Based on HTML Data

Annie Huang, Anum Damani, Rithvik Vobbilisetty, Alex Rudolph, Tyler Chia

CARPE DATA

## Introduction & Motivation

Carpe Data is an insurance technology company located in Santa Barbara, California. Carpe Data gathers and analyzes data from alternative sources such as social media platforms and web content in order to provide insurance companies with data solutions.

Our capstone team project focuses on building machine learning classification models in Python that predict and flag whether or not a web page contains evidence about a fraud claim and provide information about the specific type of activity that is present.

Insurance companies typically receive a significant amount of claims each year. By automatically flagging and classifying web pages that have information potentially relevant to the claims, our project would help to significantly reduce the amount of manual inspection required for potential cases of insurance fraud.

## Data

Our dataset provided by the Carpe Data team contained 42,485 observations representing web pages potentially containing claimant activity information. Each observation contained HTML data from the webpage and a label with the type of activity (if any):

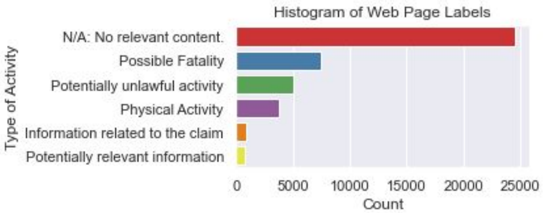


Fig. 1: Count of web page labels indicating type of claimant activity information.

## Exploratory Analysis

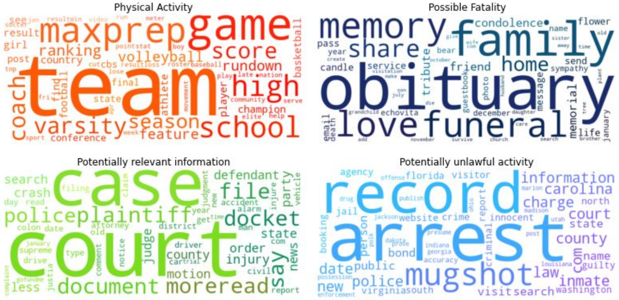
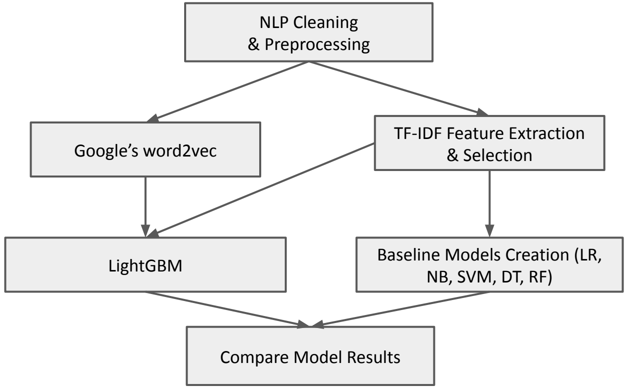


Fig. 2: Top 55 tokens using our naive bayes model for the selected labels.

## Methodology



NLP Cleaning & Preprocessing

Google's word2vec

TF-IDF Feature Extraction & Selection

LightGBM

Baseline Models Creation (LR, NB, SVM, DT, RF)

Compare Model Results

## Final Results

| Model | Weighted Avg. Precision |
|---|---|
| SVM | 85% |
| Naive Bayes | 84% |
| TF-IDF LightGBM (No Feature Selection) | 84% |
| TF-IDF LightGBM (With Feature Selection) | 81% |
| Word2Vec LightGBM | 79% |

## Conclusion

Ultimately, our best performing model was our SVM classifier with a weighted average precision of 85%. Although the other models came close in total accuracy, it seemed that our SVM classifier performed significantly better in detecting the labels 'Information related to the claim' and 'Physical activity' compared to the others.

Across all our models, the category 'Information related to the claim' was the most difficult for our models to predict, likely due to the broader range of content and language used in this label.

In the future, we'd like to improve our SVM classifier accuracy by implementing bigrams and more feature selection into our model.

## References & Acknowledgments

Sponsors: Crystal Zhang and Kevin Neal
Mentor: Joshua Bang

UC SANTA BARBARA