

---

# Classification Modelling

Amanda Khairunnisa

---

---

# Problem Statement

**Contact form**

Please fill in your information and we'll be sending your order in no time.


Your Name

Your Email

Phone\*

Message Subject\*

Message\*

Verification\* ☐ I'm not a robot 

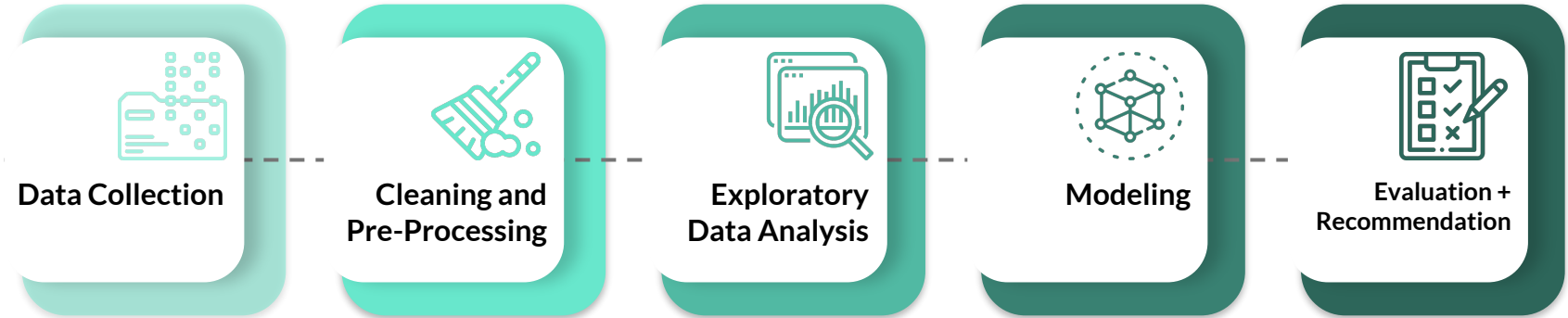
**SUBMIT FORM**

A legal firm is looking to create a contact form for their pro-bono legal clinic's website.

I have been hired to create a machine learning classification model for this new feature that would help filter out non-legal inquiries.

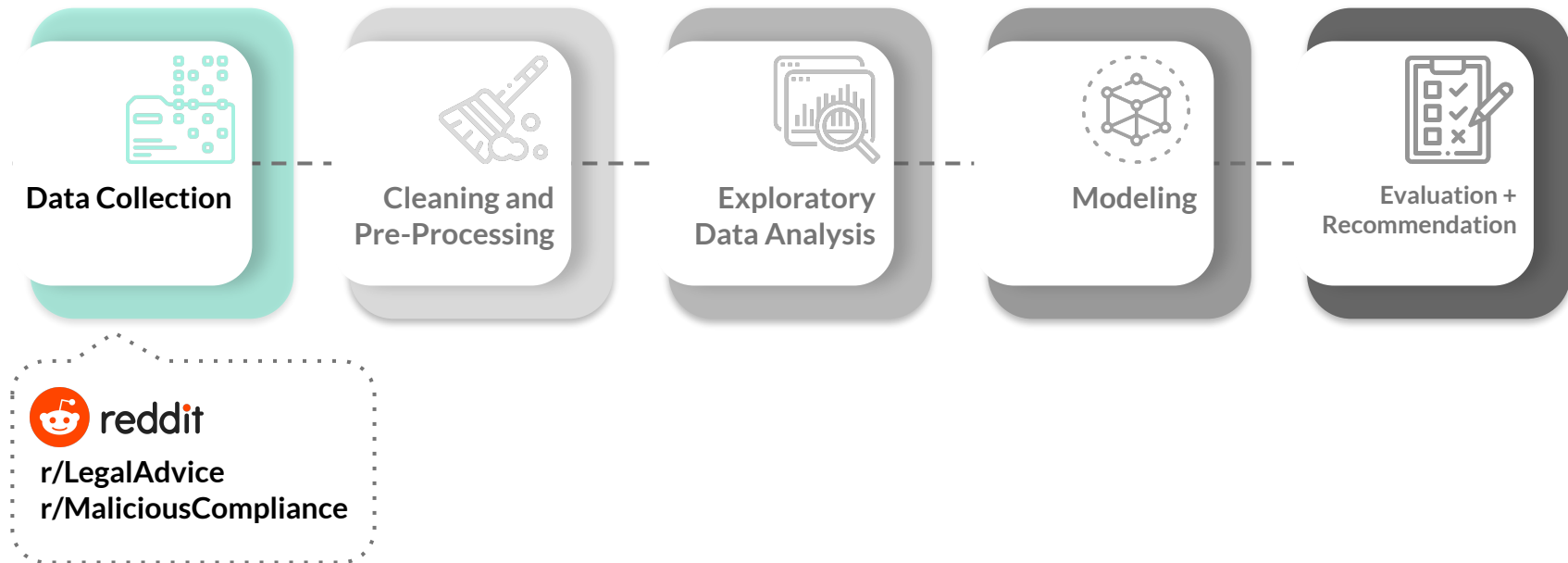
---

# Methodology



---

# Data Collection




---

# Cleaning and Pre-Processing



- Nulls
- Replacement of unusual text / characters
- Feature engineering

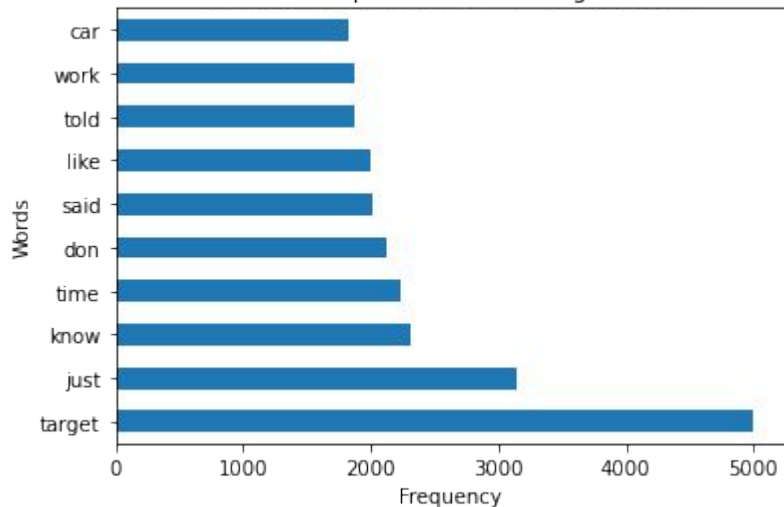
A teal square with a dashed border, centered on a white background. The text "Exploratory Data Analysis" is written in white, bold, sans-serif font, centered within the square. Two horizontal white lines are positioned above and below the text.

# Exploratory Data Analysis

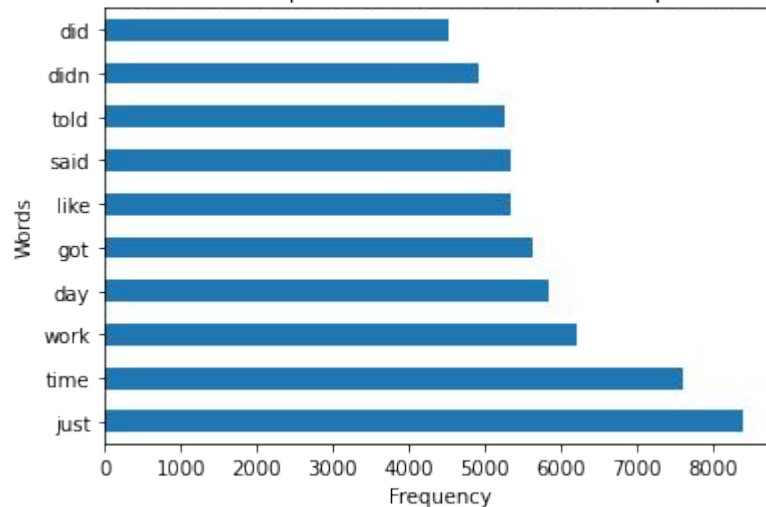
---

# Word Frequency

10 Most Frequent Words for Legal Advice



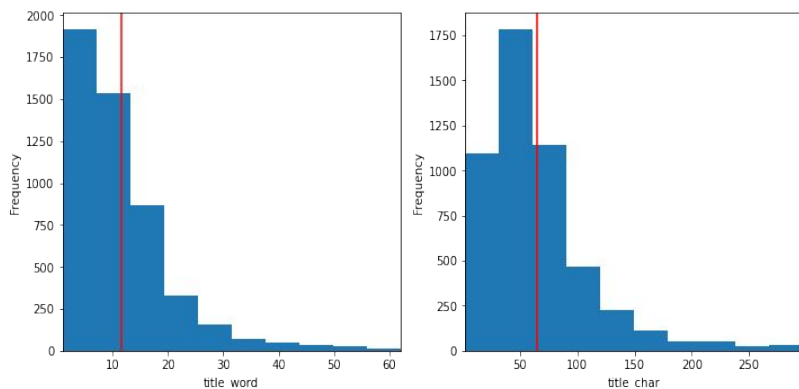
10 Most Frequent Words for Malicious Compliance



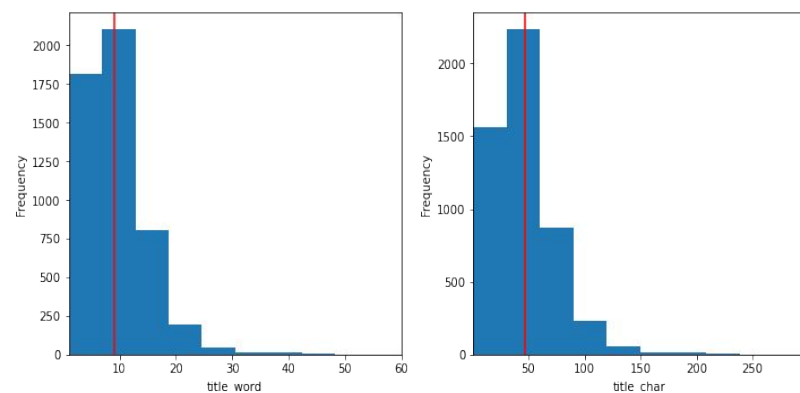
---

# Title Character and Word Distribution

## Legal Advice



## Malicious Compliance

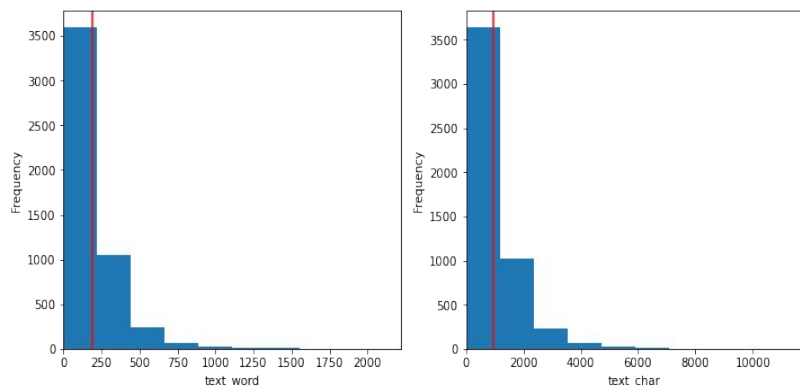




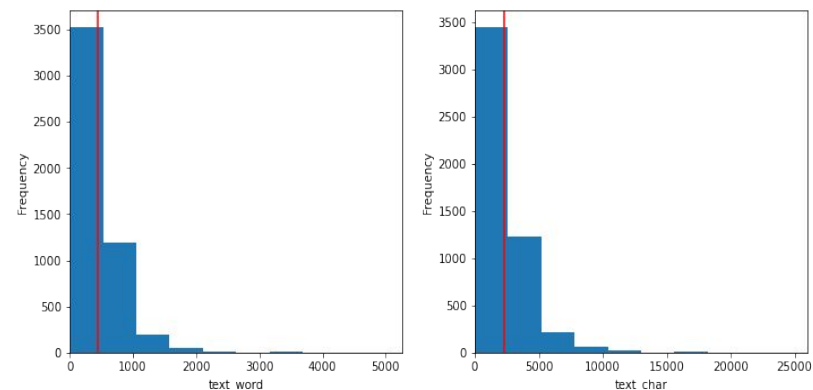
---

# Text Character and Word Distribution

## Legal Advice

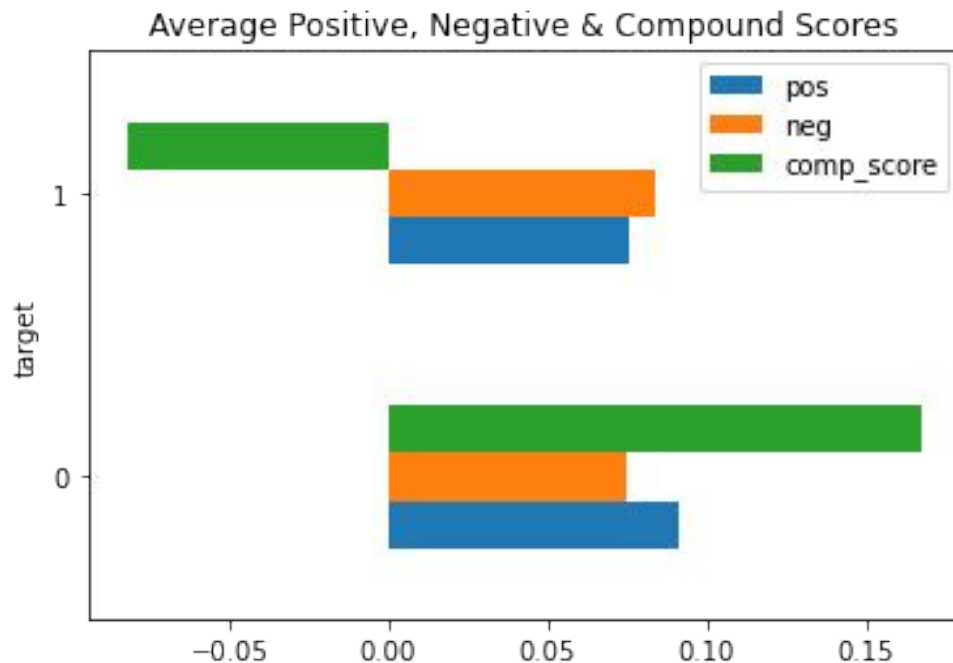


## Malicious Compliance

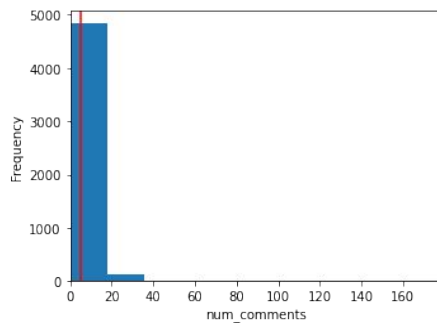


---

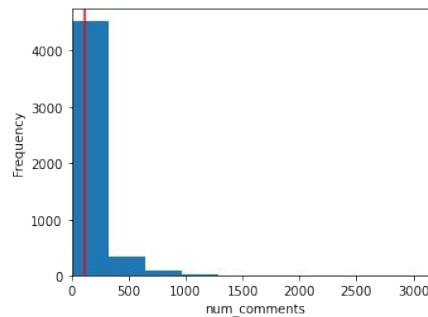
# Sentiment Analysis



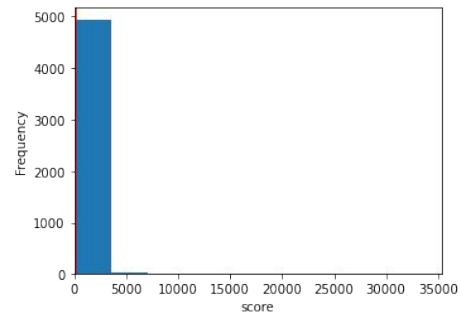
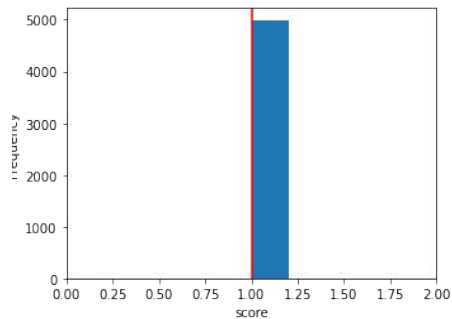
# Score and Comment Distribution



Legal Advice

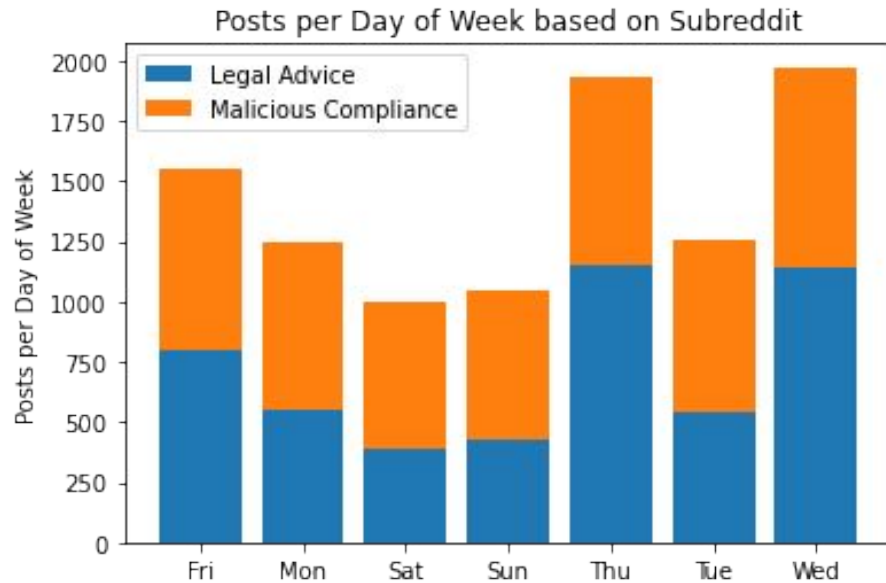


Malicious Compliance



---

# Time



---

# Modeling



- **Logistic Regression**
- **Decision Tree Classifier**
- **Random Forest Classifier**
- **Extra Trees Classifier**
- **Support Vector Machines**
- **Naive Bayes**



## **Evaluation and Recommendation**

---

# Key Performance Indicators

Baseline: 50%

Metrics	Logistic Regression	Decision Tree Classifier	Random Forest Classifier	Extra Trees Classifier	Support Vector Machines	Naive Bayes
Train Accuracy	99.9%	91.1%	94.0%	92.6%	85.1%	68.2%
Test Accuracy	96.0%	89.7%	89.8%	90.0%	85.6%	67.4%
Test Sensitivity	97.7%	92.5%	94.2%	96.0%	96.9%	99.6%
Test Specificity	94.4%	87.0%	85.4%	84.0%	74.3%	35.3%
Test ROC AUC Score	99.1%	93.0%	96.4%	96.5%	96.5%	70.0%

---

---

# Key Performance Indicators

Baseline: 50%

Metrics	Logistic Regression	Decision Tree Classifier	Random Forest Classifier	Extra Trees Classifier	Support Vector Machines	Naive Bayes
Train Accuracy	99.9%	91.1%	94.0%	92.6%	85.1%	68.2%
Test Accuracy	96.0%	89.7%	89.8%	90.0%	85.6%	67.4%
Test Sensitivity	97.7%	92.5%	94.2%	96.0%	96.9%	99.6%
Test Specificity	94.4%	87.0%	85.4%	84.0%	74.3%	35.3%
Test ROC AUC Score	99.1%	93.0%	96.4%	96.5%	96.5%	70.0%

---