CONSTRUCTION ACCIDENT REPORT CLASSIFICATION

Andrea Vagnoli
Università di Pisa
A. Y. 2024 - 2025

INTRODUCTION

According to the ILO, about 2.78 million workers die annually from occupational accidents, with one in six occurring in the construction industry.

Following an accident, detailed **reports** are usually compiled, which include also unstructured narrative data (e.g., descriptions and summaries of the event). Their unstructured nature poses considerable challenges for analysis and knowledge extraction.

The objective of this project is to develop a **classification model** capable of assigning **construction accident reports** to their **correct category**.

<table>
<tr><td colspan="6">Accident: 114404.015 - Employee Falls From Roof And Dies From Multiple Injuries</td></tr>
<tr><td>Open Date</td><td colspan="2">Establishment Name</td><td>End-Use</td><td>Project Type</td><td>…</td></tr>
<tr><td>03/13/2019</td><td colspan="2">Hough Roofing, Inc.</td><td>Commercial building</td><td>Maintenance or repair</td><td>…</td></tr>
<tr><td colspan="6">At 4:00 p.m. on March 12, 2019, Employee #1, employed by a roofing company, was engaged in roofing work at a two-story commercial building... It began to rain slightly. Employee #1 fell, a fall height of 23.5 feet… Employee #1 died later that night from his injuries.</td></tr>
<tr><td colspan="6">Keywords: roofer, fall, fall protection, construction, …</td></tr>
<tr><td>Employee #</td><td>Age</td><td>Sex</td><td>…</td><td>Construction</td><td>Inspection</td></tr>
<tr><td>1</td><td>52</td><td>M</td><td>…</td><td>FatCause: Fall from roof</td><td>1384743.015</td></tr>
<tr><td colspan="6">Inspection: 1384743.015 - Hough Roofing, Inc.</td></tr>
<tr><td colspan="6">Violation Items</td></tr>
<tr><td>ID</td><td>Type</td><td colspan="2">Standard</td><td>Curr$</td><td>Init$</td></tr>
<tr><td>01001</td><td>Serious</td><td colspan="2">19260501 B11</td><td>$11,934</td><td>$13,260</td></tr>
<tr><td>01002</td><td>Serious</td><td colspan="2">19260503 A01</td><td>$2,652</td><td>$5,304</td></tr>
<tr><td colspan="6">…</td></tr>
</table>

DATASET

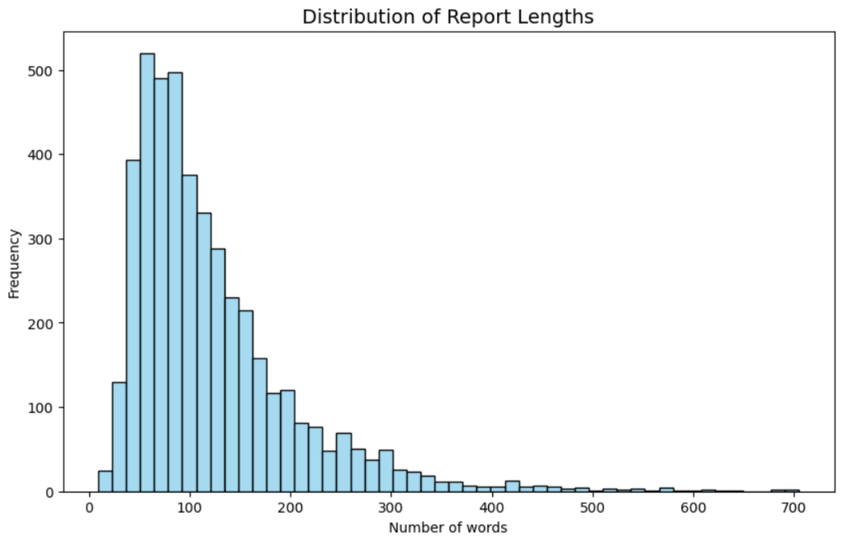
4,770 construction accident reports from the Occupational Safety and Health Administration (OSHA).

There are several fields, but we are interested in the following:

• **title and SUMMARY**: The title and the narrative text describing the details of the accident;

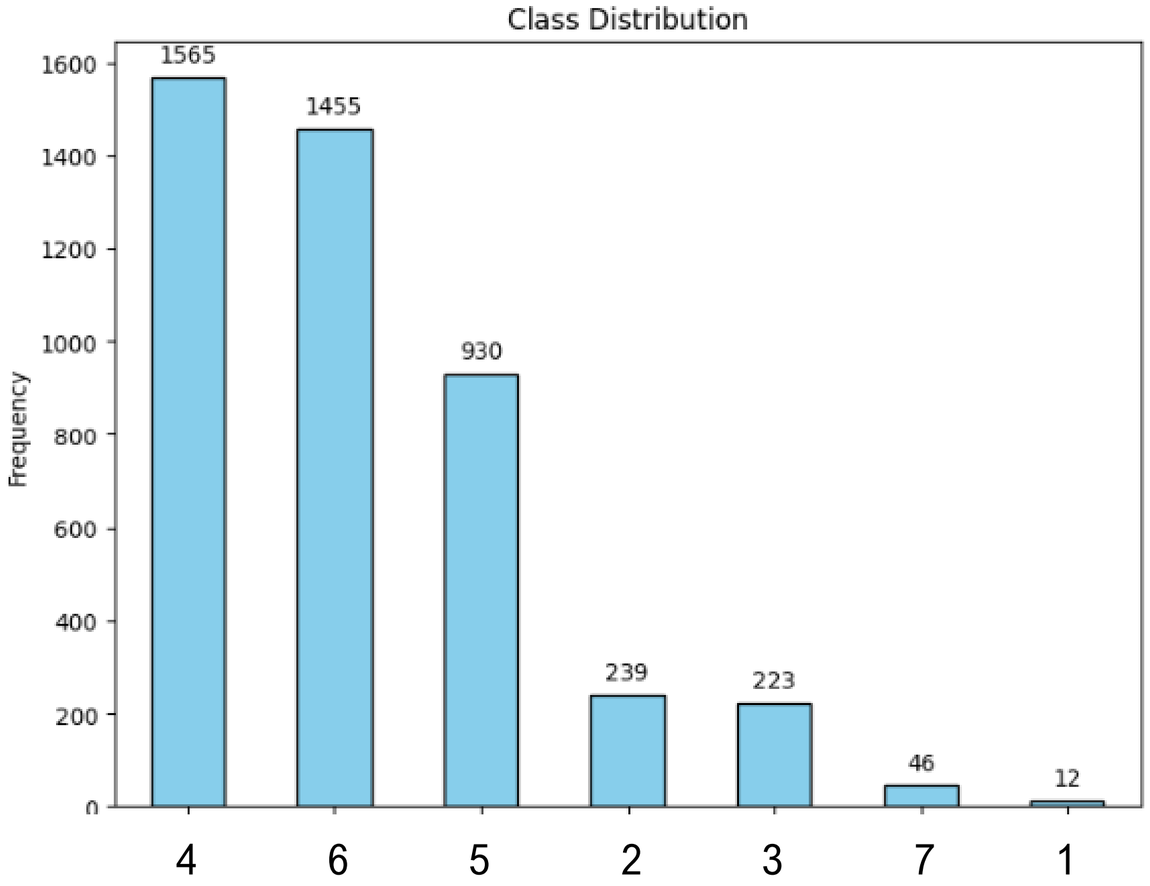
• **TaggedL1**: This is the primary label of the accident. There are seven distinct categories:

– 1 VIOLENCE AND OTHER INJURIES BY PERSONS OR ANIMALS
– 2 TRANSPORTATION INCIDENTS
– 3 FIRES AND EXPLOSIONS
– 4 FALLS, SLIPS, TRIPS
– 5 EXPOSURE TO HARMFUL SUBSTANCES OR ENVIRONMENTS
– 6 CONTACT WITH OBJECTS AND EQUIPMENT
– 7 OVEREXERTION AND BODILY REACTION



CLASS DISTRIBUTION

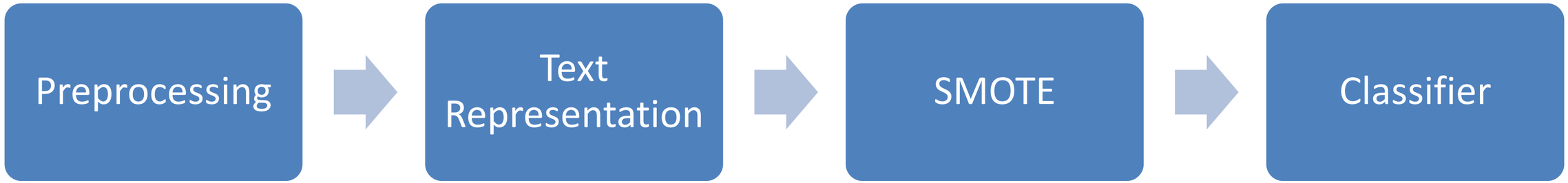
- Highly **imbalanced problem;**
- Class overlap, especially for minority classes (harder to recognize them) [1][2].



DATA PREPARATION

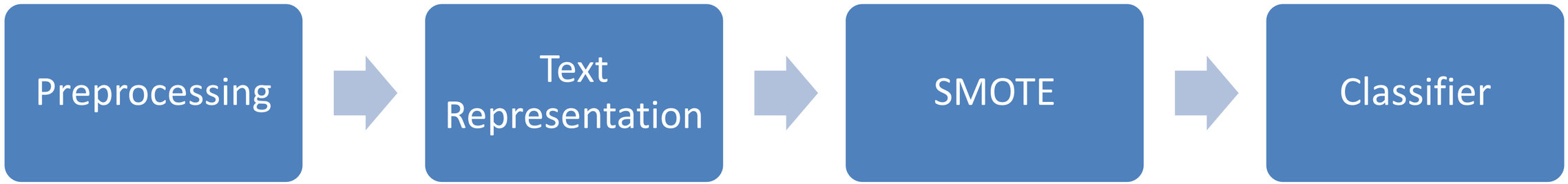
- Combined the **title** and **SUMMARY** fields to create the input texts (**X**);
- Used **TaggedL1** as the target labels (**y**);
- Performed a 70/30 train – test split.

PIPELINE BUILD



- Tokenization;
- Convert to lowercase;
- Removal of punctuation and stopwords;
- Stemming.

PIPELINE BUILD



- Word Embedding: Word2Vec
- Text Vectorizer: Weighted Class TF-IDF

TEXT REPRESENTATION

WCTF-IDF Algorithm:

1. Set `f` as the maximum total number of features.

2. Sort the classes in descending order of document frequency.

3. For each class `i`:
   - Compute the number of features to assign:
     $f_i = f \times (n_i / n)$

     where n_i = number of documents in class i,
     and n = total number of documents.

   - Fit a TF-IDF vectorizer on the documents of class i
     using f_i as the max_features parameter.

   - Pass the selected terms as stopwords to the next class
     to reduce vocabulary overlap.

4. Merge all resulting vocabularies to form the final TF-IDF vectorizer.

MODEL EVALUATION

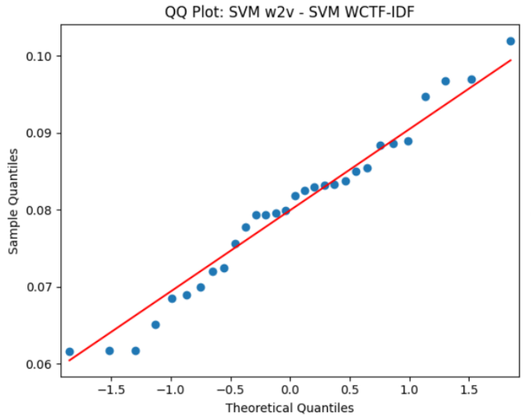
- Several different classifiers;
- **Nested 5-fold cross-validation:** (Inner and outer loops);
- **Stratified sampling** to ensure balanced classes.

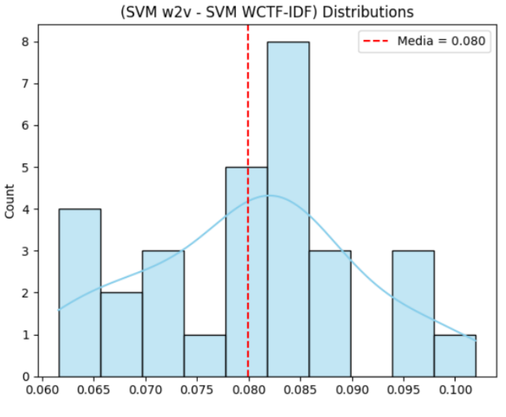
Model	WCTF-IDF		Word2Vec	
	Accuracy	Weighted F1-score	Accuracy	Weighted F1-score
Random Forest	0.883 ± 0.007	0.879 ± 0.007	0.817 ± 0.010	0.815 ± 0.012
Logistic Regression	$\mathbf{0.903 \pm 0.007}$	$\mathbf{0.903 \pm 0.007}$	0.808 ± 0.005	0.818 ± 0.003
Linear SVM	0.901 ± 0.008	0.901 ± 0.008	0.820 ± 0.009	0.826 ± 0.007
XGBoost	0.891 ± 0.006	0.889 ± 0.006	0.826 ± 0.008	0.825 ± 0.007
Bagging	0.824 ± 0.015	0.829 ± 0.012	0.786 ± 0.019	0.787 ± 0.017
Decision Tree	0.794 ± 0.014	0.798 ± 0.013	0.680 ± 0.016	0.688 ± 0.015
KNN	0.680 ± 0.007	0.705 ± 0.008	0.727 ± 0.013	0.744 ± 0.012
MultinomialNB	0.856 ± 0.017	0.856 ± 0.016	—	—

MODEL SELECTION

- SVM WCTF-IDF vs SVM word2vec (Wilcoxon Test)

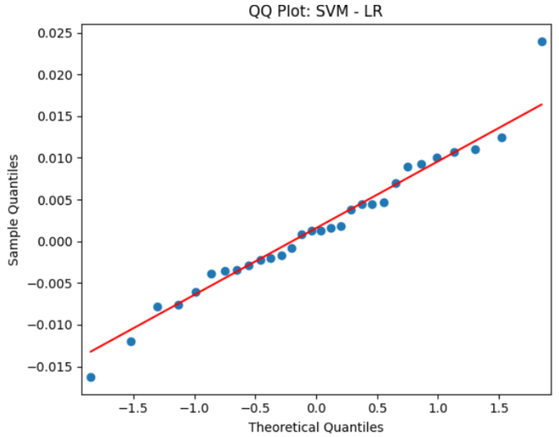
p-value = 0.0000000019 → Statistical evidence of difference.

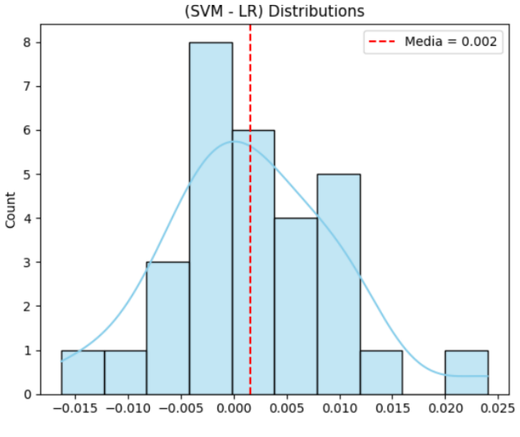




- SVM WCTF-IDF vs LR WCTF-IDF (Wilcoxon Test)

p-value = 0.36 → No statistical evidence of difference.

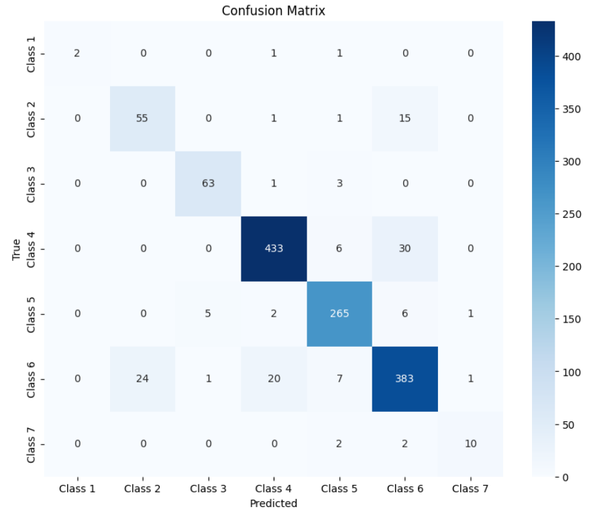




PERFORMANCE EVALUATION (Test set)

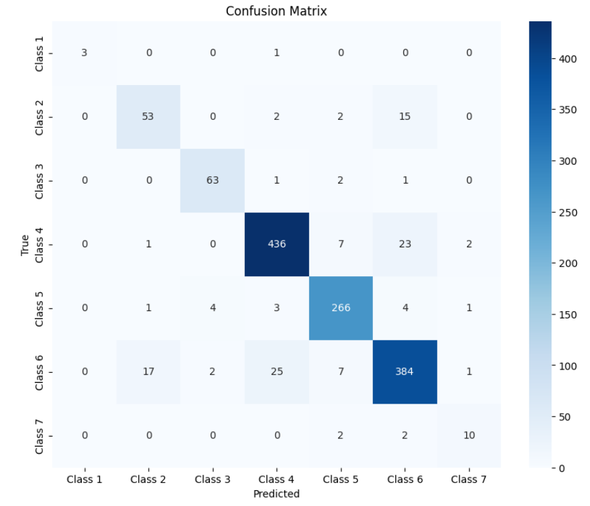
- SVM

Class	Precision	Recall	F1-Score	Support
Class 1	1.00	0.50	0.67	4
Class 2	0.70	0.76	0.73	72
Class 3	0.93	0.94	0.93	67
Class 4	0.95	0.92	0.93	469
Class 5	0.93	0.95	0.94	279
Class 6	0.88	0.88	0.88	436
Class 7	0.83	0.71	0.77	14
Accuracy		0.90		1341
Macro Average	0.89	0.81	0.83	1341
Weighted Average	0.90	0.90	0.90	1341



- Logistic Regression

Class	Precision	Recall	F1-Score	Support
Class 1	1.00	0.75	0.86	4
Class 2	0.74	0.74	0.74	72
Class 3	0.91	0.94	0.93	67
Class 4	0.93	0.93	0.93	469
Class 5	0.93	0.95	0.94	279
Class 6	0.90	0.88	0.89	436
Class 7	0.71	0.71	0.71	14
Accuracy		0.91		1341
Macro avg	0.87	0.84	0.86	1341
Weighted avg	0.91	0.91	0.91	1341



COMPARISON WITH OTHER STUDIES

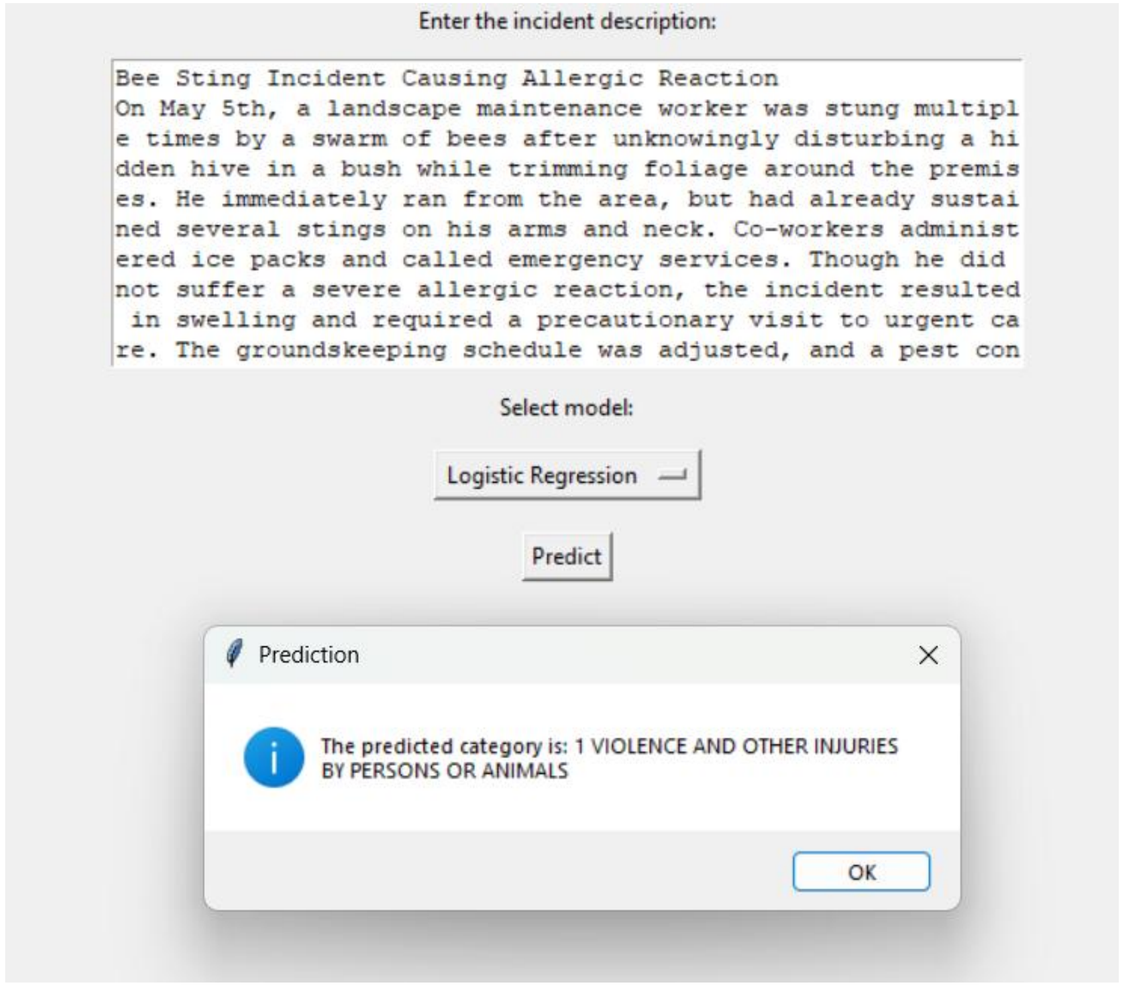
Reference paper results[2]:

- Best model: **SVM** with **8,423 features** (vs. **1,000 features** in our case);

- Overall **Accuracy** and **Weighted F1-score** around **0.91** (same as our best model);

- Class-wise F1-score for the most imbalanced classes (1 and 7):
  Paper: 0.40 and 0.62
  Our best results: **0.86** and **0.77**

INTERFACE


Enter the incident description:
Bee Sting Incident Causing Allergic Reaction
On May 5th, a landscape maintenance worker was stung multipl
e times by a swarm of bees after unknowingly disturbing a hi
dden hive in a bush while trimming foliage around the premis
es. He immediately ran from the area, but had already sustai
ned several stings on his arms and neck. Co-workers administ
ered ice packs and called emergency services. Though he did
not suffer a severe allergic reaction, the incident resulted
in swelling and required a precautionary visit to urgent ca
re. The groundskeeping schedule was adjusted, and a pest con
Select model:
Logistic Regression
Predict
Prediction
The predicted category is: 1 VIOLENCE AND OTHER INJURIES
BY PERSONS OR ANIMALS
OK

REFERENCES

[1] Cheng, M. Y., Kusoemo, D., & Gosno, R. A. (2020). Text mining-based construction site accident classification using hybrid supervised machine learning. Automation in Construction, 118, 103265.

[2] Qiao, J., Wang, C., Guan, S., & Liu, S. (2022). Construction-accident narrative classification using shallow and deep learning. Journal of Construction Engineering and Management, 148(9).

[3] Deepwiz AI. (2023). How to correctly use TF-IDF with imbalanced data. Retrieved from https://www.deepwizai.com/projects/ how-to-correctly-use-tf-idf-with-imbalanced-data

THANK YOU FOR YOUR ATTENTION