



# 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. 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**.

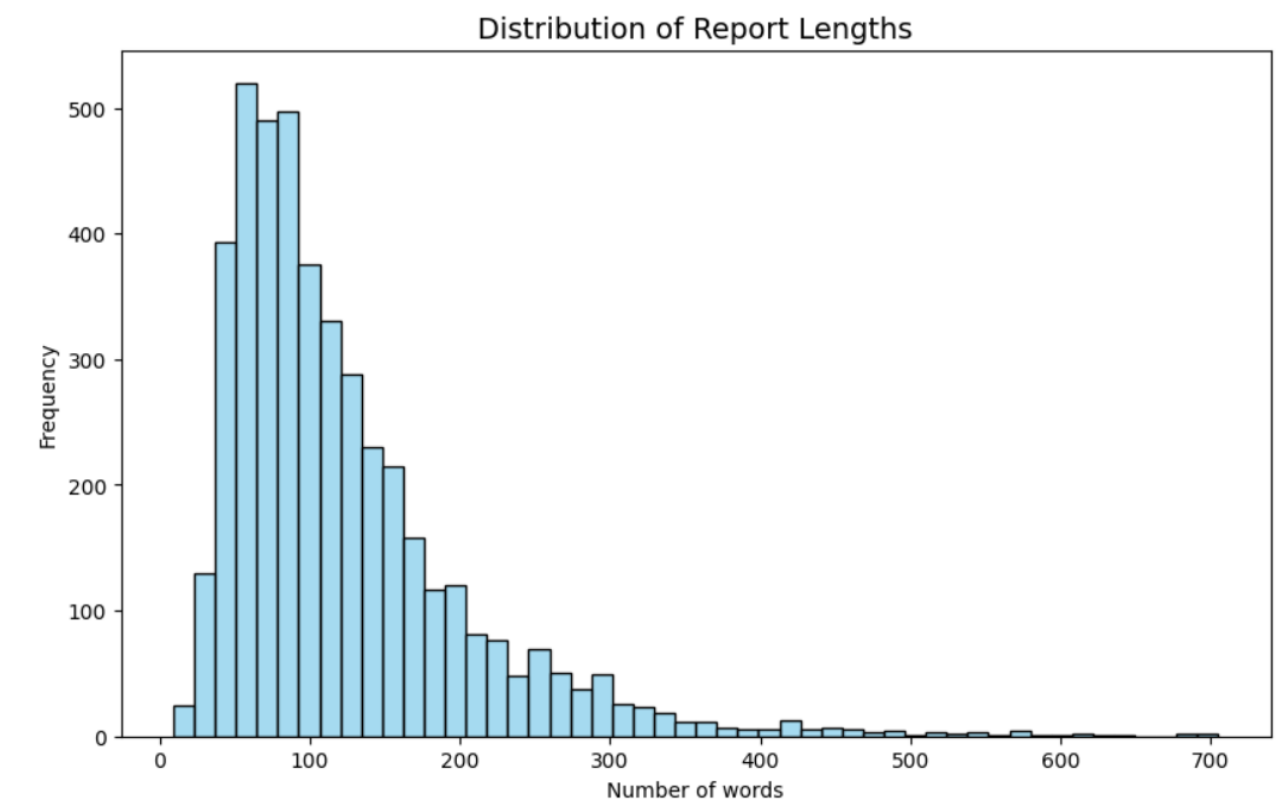
Accident: 114404.015 - Employee Falls From Roof And Dies From Multiple Injuries					
Open Date	Establishment Name		End-Use	Project Type	...
03/13/2019	Hough Roofing, Inc.		Commercial building	Maintenance or repair	...
At 4:00 p.m. on March 12, 2019, Employee #1, employed by a roofing company, was engaged 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.					
Keywords: roofer, fall, fall protection, construction, ...					
Employee #	Age	Sex	...	Construction	Inspection
1	52	M	...	FatCause: Fall from roof	1384743.015
Inspection: 1384743.015 - Hough Roofing, Inc.					
Violation Items					
ID	Type	Standard		Curr\$	Init\$
01001	Serious	19260501 B11		\$11,934	\$13,260
01002	Serious	19260503 A01		\$2,652	\$5,304
...					

# DATASET

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

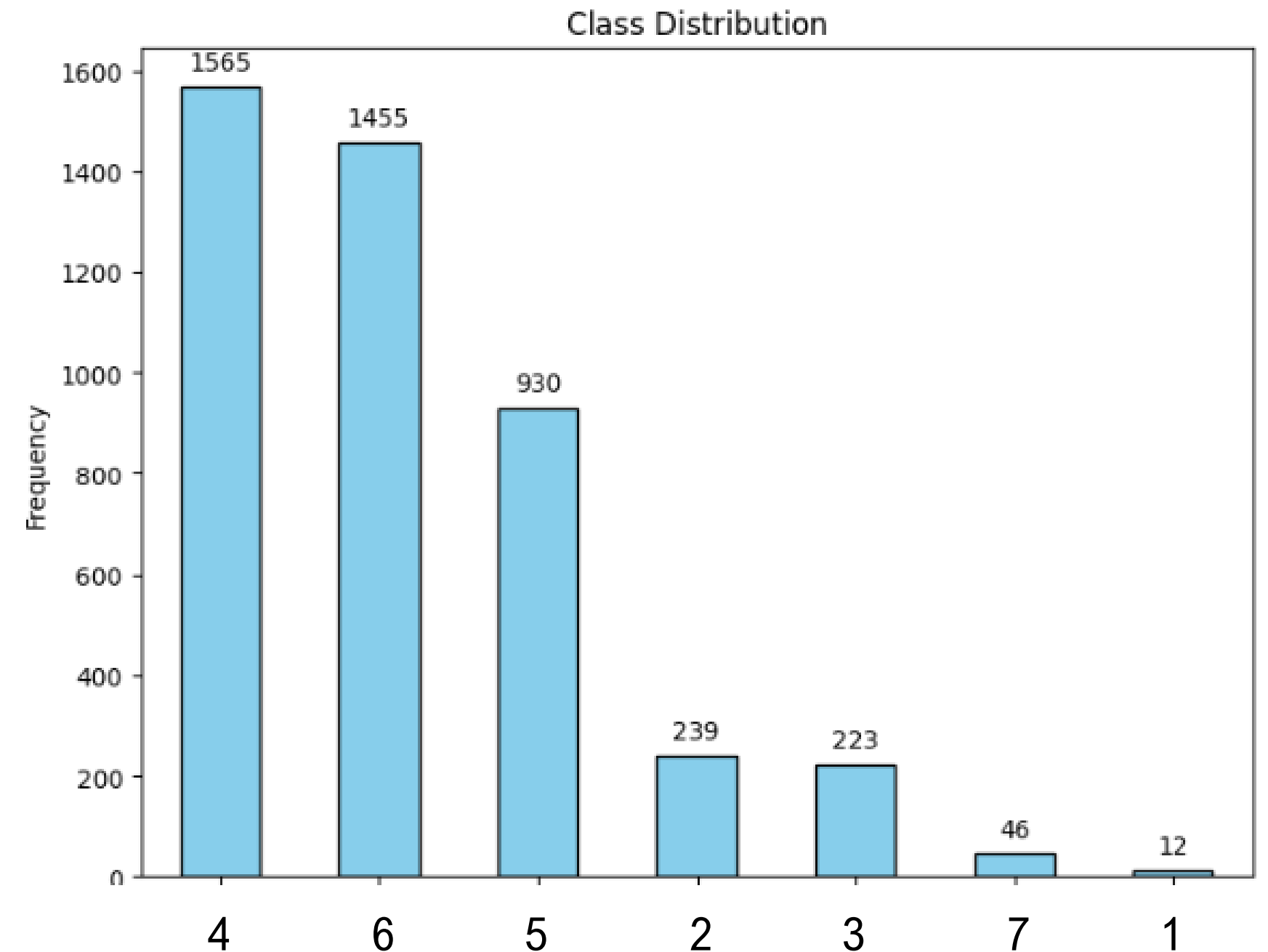
There are several fields, but we are interested in the following:

- **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].



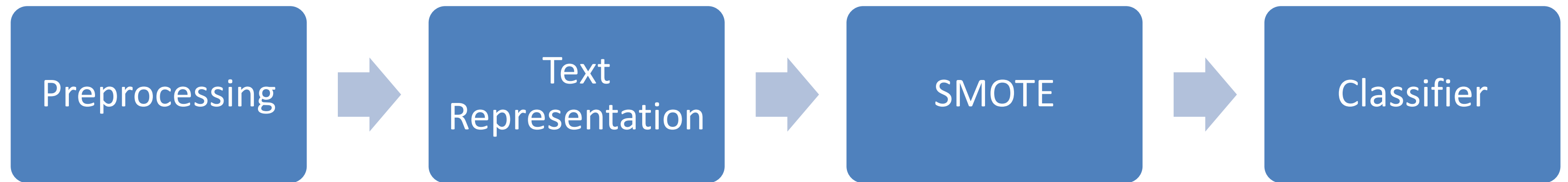
# DATA PREPARATION

---

- 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

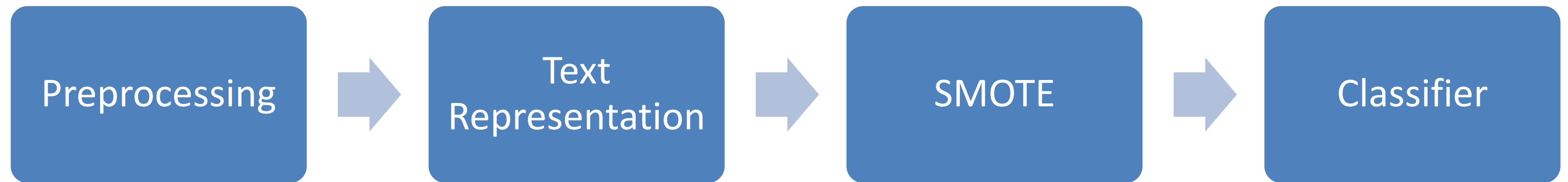
---



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

# PIPELINE BUILD

---



- Word Embedding: Word2Vec
- Text Vectorizer: Weighted Class TF-IDF[3]

# TEXT REPRESENTATION

WCTF-IDF Algorithm (Weighted-Class TF-IDF)

1. Set `f` as the maximum total number of features.
2. Sort classes in descending order by document frequency.
3. For each class `i`:
  - Compute number of features to assign:  
$$f_i = f \times (n_i / n)$$
where:
    - $n_i$  = number of documents in class `i`,
    - $n$  = total number of documents.
  - Fit a TF-IDF vectorizer on class `i` documents using `fi` as `max_features`.
  - Save the vocabulary extracted for class `i`.
  - Pass selected terms as stopwords to the next class to reduce overlap.
4. Merge all class-specific vocabularies into a unified set.
5. Fit a final TF-IDF vectorizer on the entire dataset using the merged vocabulary.



# 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	$0.883 \pm 0.007$	$0.879 \pm 0.007$	$0.817 \pm 0.010$	$0.815 \pm 0.012$
Logistic Regression	<b><math>0.903 \pm 0.007</math></b>	<b><math>0.903 \pm 0.007</math></b>	$0.808 \pm 0.005$	$0.818 \pm 0.003$
Linear SVM	$0.901 \pm 0.008$	$0.901 \pm 0.008$	$0.820 \pm 0.009$	$0.826 \pm 0.007$
XGBoost	$0.891 \pm 0.006$	$0.889 \pm 0.006$	$0.826 \pm 0.008$	$0.825 \pm 0.007$
Bagging	$0.824 \pm 0.015$	$0.829 \pm 0.012$	$0.786 \pm 0.019$	$0.787 \pm 0.017$
Decision Tree	$0.794 \pm 0.014$	$0.798 \pm 0.013$	$0.680 \pm 0.016$	$0.688 \pm 0.015$
KNN	$0.680 \pm 0.007$	$0.705 \pm 0.008$	$0.727 \pm 0.013$	$0.744 \pm 0.012$
MultinomialNB	$0.856 \pm 0.017$	$0.856 \pm 0.016$	—	—

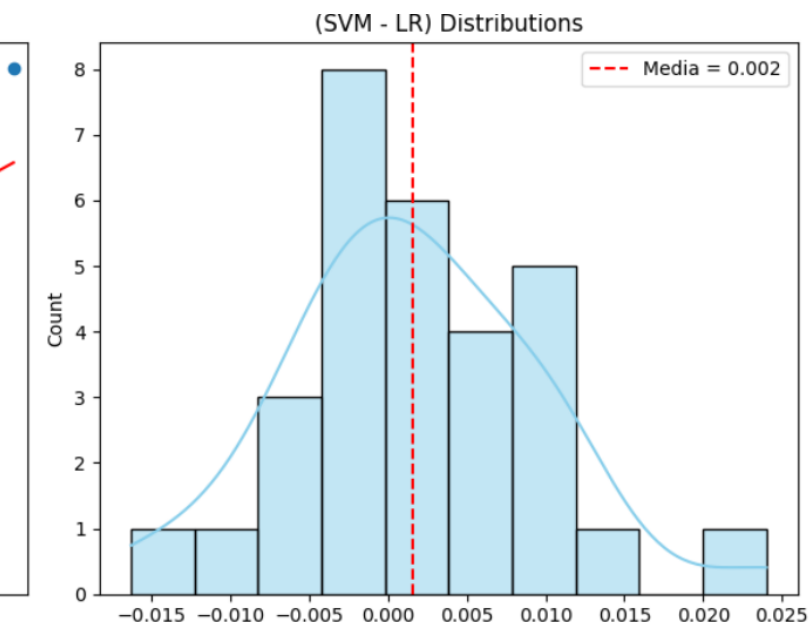
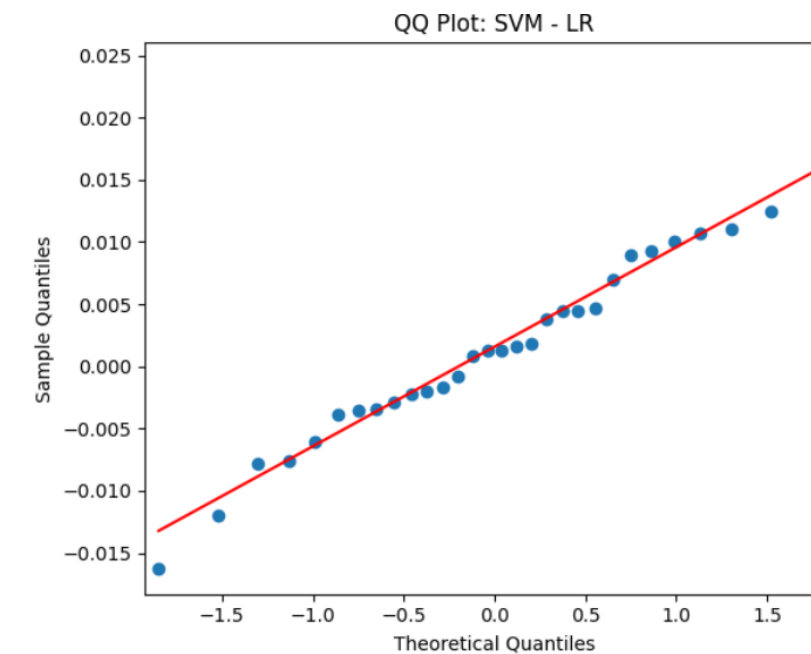
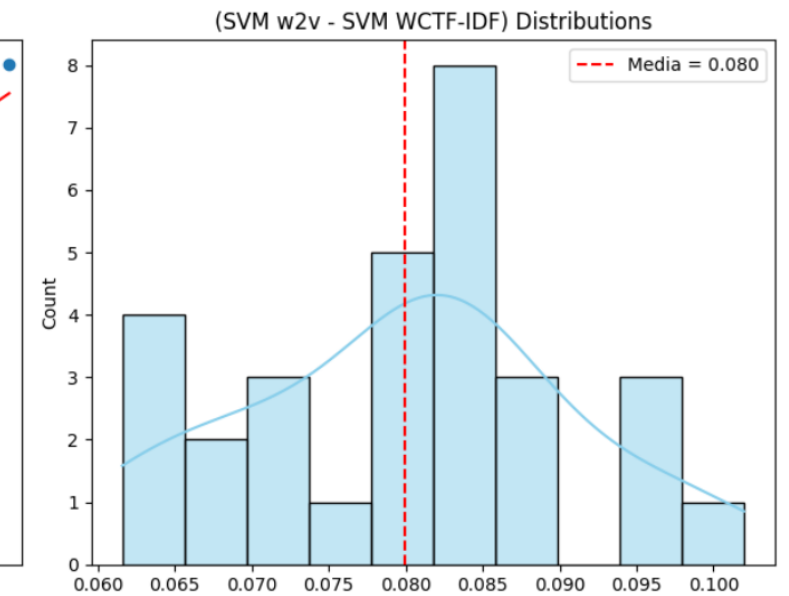
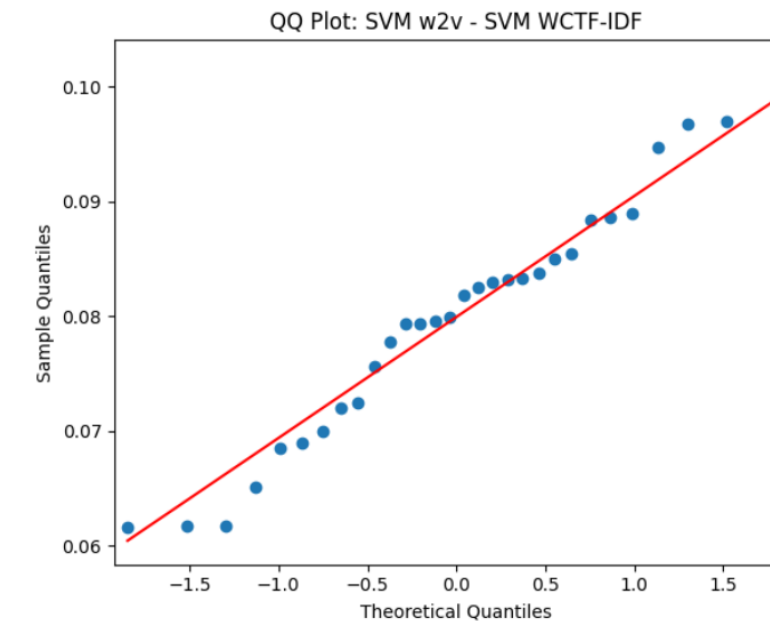
# MODEL SELECTION

- 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)

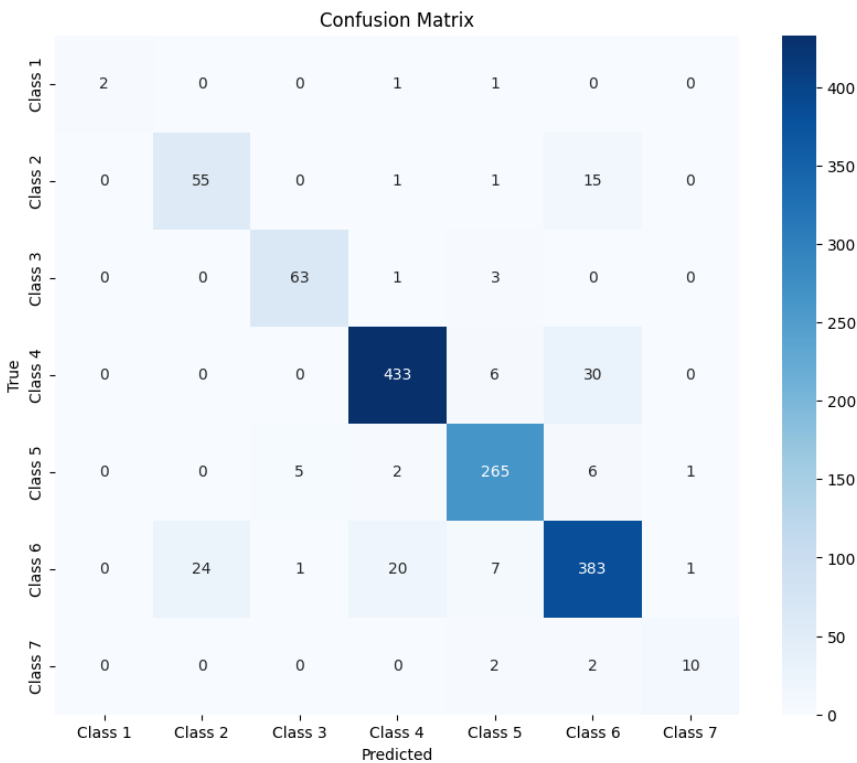
p-value = 0.36 → No statistical evidence of difference.



# PERFORMANCE EVALUATION (Test set)

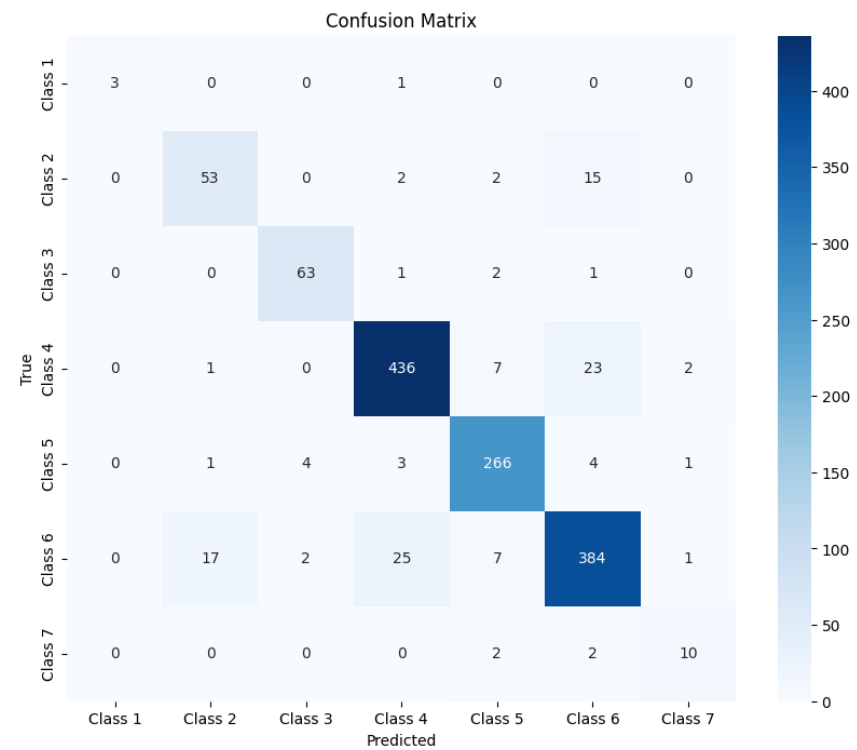
• SVM

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 multiple times by a swarm of bees after unknowingly disturbing a hidden hive in a bush while trimming foliage around the premises. He immediately ran from the area, but had already sustained several stings on his arms and neck. Co-workers administered 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 care. 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**

