Department of Civil Engineering
Indian Institute of Technology, Madras
August 2020 – June 2021

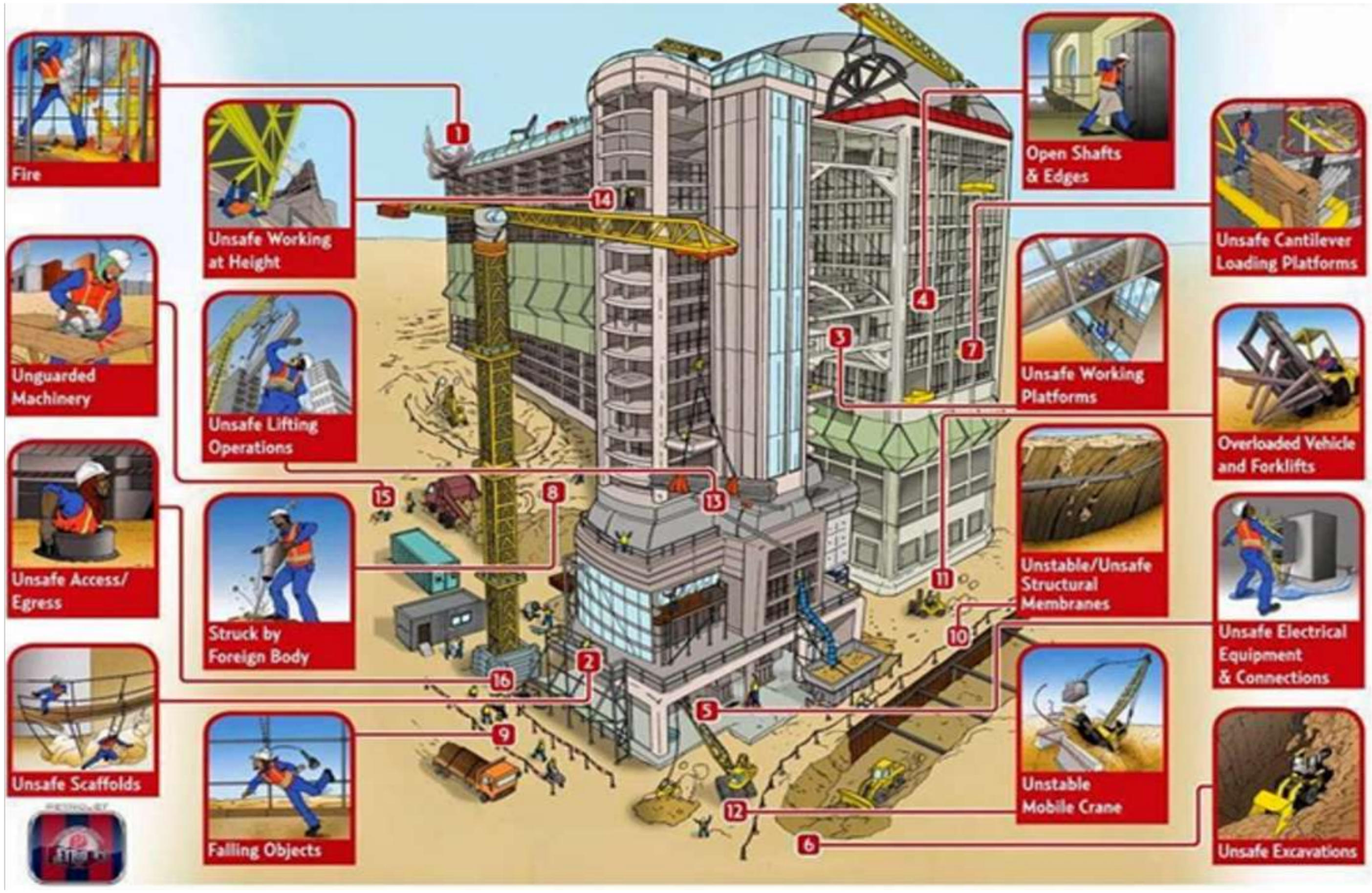
BACHELOR'S PROJECT REVIEW

Analyzing Safety Reports using ML/NLP

Guided by
Dr. Nikhil Bugalia
Dr. Ashwin Mahalingam

Jai Kedia
CE17B037

Introduction



- ➔ Construction continues to be among one of the most dangerous occupations around the world.
- ➔ With such a diverse set of activities involved, it is essential to pay due attention to safety.

Background

➔ Safety observation Reports (SORs) collected at site can be helpful to understand the causes of accidents and prevent them.

➔ With the massive data collected every day, it requires lot of effort and time to analyse the reports.

➔ Hence there is a need to use Machine learning to analyse SORs to predict the accidents and analyse their causes.

➔ An accident can either lead to injury or cause damage or it might be just a near miss.

➔ There can be multiple causes leading to an accident which can be either the ignorance of management or the worker.

Health and safety incident report form

Workers Health & Safety Centre

The incident

Reported by | Department

Email | Phone | Ext

Date of occurrence | Time

Exact location

Accident ☐ Incident ☐ Near miss ☐ Violence ☐ Ill health ☐ Safety ☐

What happened? Report any details that may have contributed to the incident (i.e., poor lighting). Use additional paper as necessary and attach to form.

Describe the outcome: harm/health effects/damage.

Describe corrective measures taken to address immediate hazards related to incident.

Research Gaps

➔ A majority of studies are done on structured data obtained from various organisations.

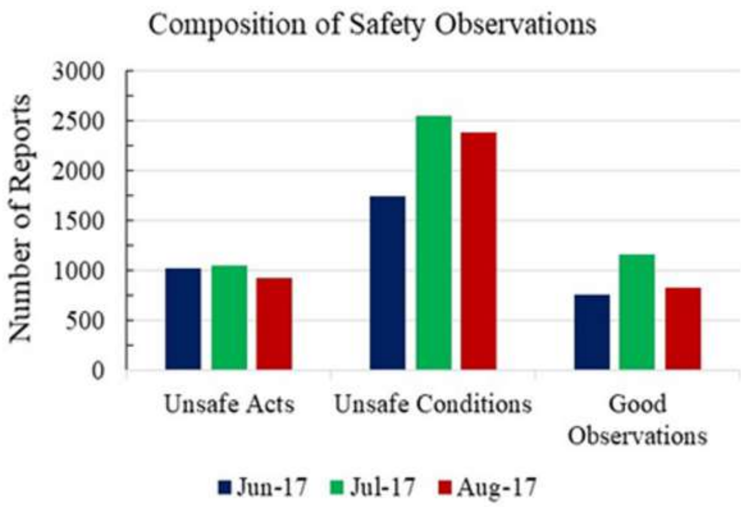
➔ Emphasis of previous studies was given more on analysing causes and outcomes on accident reports leading to injury or damage using machine learning.

Objective

➔ Aim is to develop an efficient strategy for analysis of safety observations obtained directly from the construction sites.

➔ To use different Machine learning and text mining based approaches to reduce the need for manual analysis of the high voluminous reports.

Data (12428 observations)



Unsafe Act - Activity by workers which are not as per the prescribed safety standard or practice and which can cause or likely to cause accidents

Unsafe Condition - Any condition or situation (electrical, chemical, biological, physical, mechanical, management and environmental) which increases the risks and dangers of accidents can be called as unsafe conditions

Good Observations- Some positive act/condition

The data is collected from a large-scale construction site on a natural gas plant in Kuwait.

4000 observations/month, Average 4 Million man-hours/month, About 21,000 workers

At this site, the workers from several non-English speaking countries gathered and, have provided SOs and their classification (UA, UC or GO) in a textual format

Data Quality Examples

	Real data “before spell check”	Real data “After spell check”	Database data
Total number of words	109657	109657	67974
Percentage of unknown words	7.26	2.24	1.87
Average word count	8.8	8.8	50+

Very less contextual information such as the task assigned, instructions provided, work being done etc.

Unstructured English, Grammar patterns, spelling mistakes

Data Quality Examples

OBSERVATION (before)	OBSERVATION (Ater Cleaning)	Worker label	Predicted label	Category
Observe seobon crew maintain and following swp of excavation safety.	observe seobon crew maintain following swap excavation safety	UC	GO	Mislabelled
As heat is raising there is a need of supply of ors to the workers	heat raising need supply worker	GO	UC	Mislabelled

A lot of context specific jargon – swp (safe working procedure?), ors (Oral Rehydration Salts?)

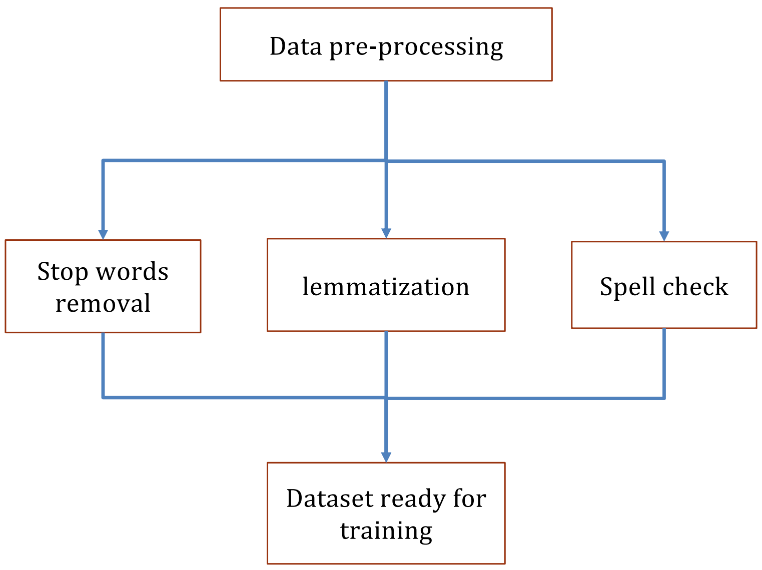
Potentially mislabelled by the workers –

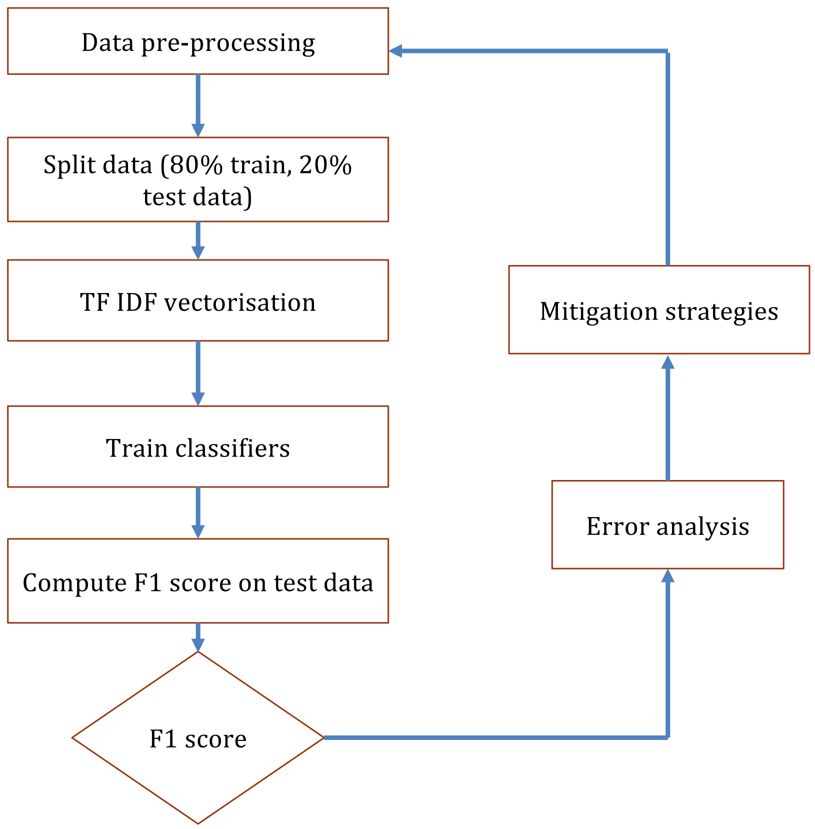
Biasness to report a negative comment about your friends

Management's promotion to report positively to create a safe environment

Genuine problems in distinguishing UA/UC etc. (Difficulty for the worker to observe the full context)

Methodology





The classifier input is the textual description, and the output is the classified label

The calculation here is fully automated, does not require manual inputs to process

Measure Of Accuracy

Actual	Predicted	
	True	***False***
True	True Positives	False Negatives
False	False Positives	True Negatives

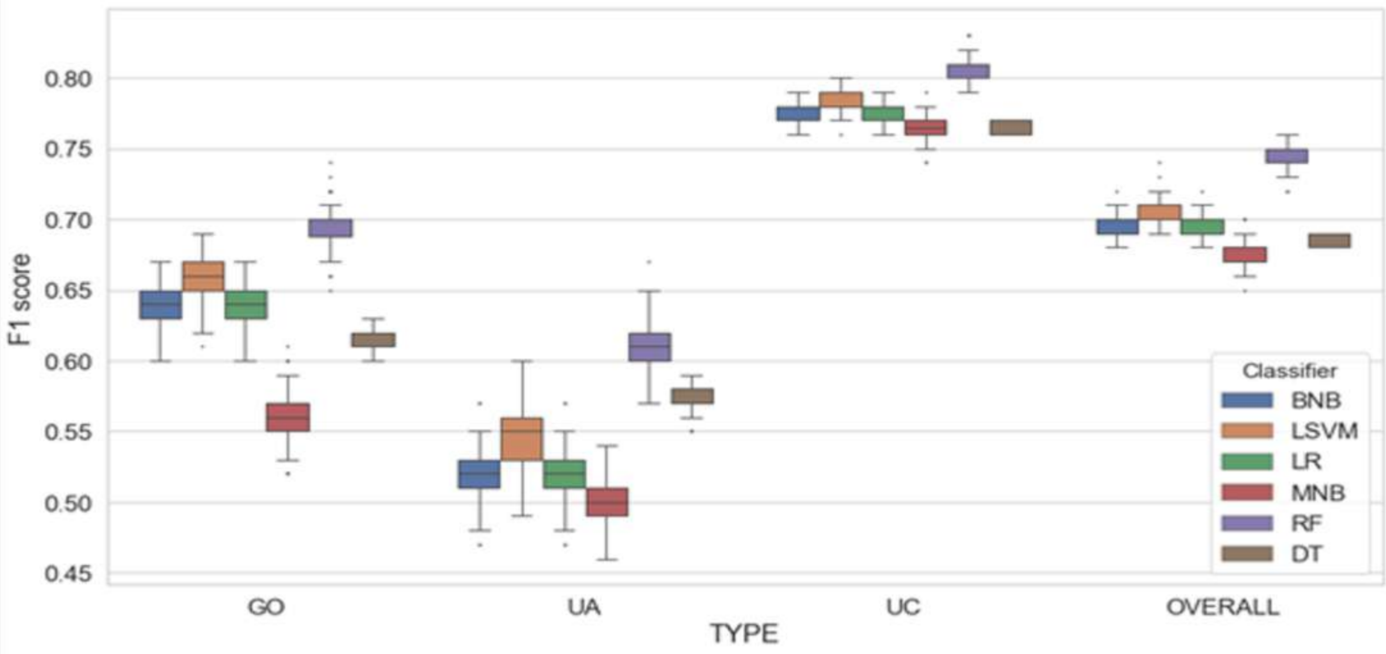
Precision – TP/(TP+FP)

Recall – TP/(TP+FN)

F-Score – Harmonic Mean (Precision, Recall)

Roughly speaking, F-Score gives an idea of what % of the observations are classified correctly

Preliminary Results



BNB – Bernoulli naïve Bayes
L-SVM – Support vector machine (linear)
LR – Logistic regression
MNB – Multinomial Naïve Bayes
RF – Random Forest
DT – Decision Tree

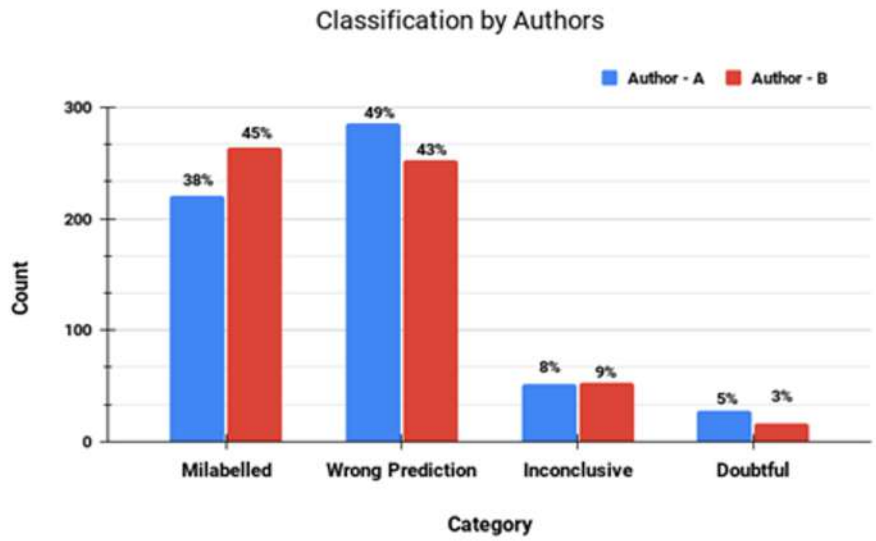
The six algorithms were run on 100 random stratified samples, to calculate the F-1 scores of the categories UA, UC, GO.

Classification accuracy – **UC** *(high proportion in original data)* **>GO** *(positive features that are repetitive)* >**UA**

Overall, RF classifier gives an average prediction accuracy of approximately 80% consistent with previous literature (Goh and Ubeynarayana, 2017; Marucci-Wellman et al., 2017).

Error Analysis

OBSERVATION (before)	OBSERVATION (Ater Cleaning)	Worker label	Predicted label	Category
No sign board at hydro test area.	sign board hydro test area	UC	GO	Wrong Prediction
safe access siganges to be provided in col 9 of U-151	safe access siganges provided col u	UC	GO	Wrong Prediction
Observe seobon crew maintain and following swp of excavation safety.	observe seobon crew maintain following swap excavation safety	UC	GO	Mislabelled
As heat is raising there is a need of supply of ors to the workers	heat raising need supply worker	GO	UC	Mislabelled



The testing data along with assigned labelled and predicted label were categorised in to 4 types by two authors.

A major proportion i.e., approximately 46% of the errors were classified to be “Wrong Prediction”,

The issues relating to the exclusion of certain important words, as part of the stop-word removal process. (Example highlighted in Yellow)

Mitigation (Stop-Word)

Stop Words- Words removed from the data during pre-processing.

Usually, a standard list is used.

However, we needed to remove certain stop-words

The stop words excluded in the study

“no", “wouldn’t”, "during", “didn’t”, “not”, "above", "below", "did", "shouldn’t”, "before", "after", "had", "have", "will", ”against

Not 100% effective – ML models cannot handle too many unique words

Observation	Observation (Processed)	Category Labelled	Predicted
NO PROPER COLOUR CODE SOME TOOLS	proper colour code tool	UC	GO
When lifting in progress there is no proper sign and proper barication.	lifting progress proper sign proper barication	UA	GO
When workers work, they didn't housekeeping	worker work housekeeping	UC	GO
Pipe rack 9 found one helper w/out gloves using welding cable pulling	pipe rack found one helper w glove using welding cable pulling	UA	GO

Mitigation (Mislabel Correction)

Three Authors

Inter-rater reliability score, 600 observations - Kappa (0.65, Good Agreement)

Then, individual relabelling, and discussions.

Final, mislabel correction for all the GOs

Challenges in Mislabel correction

- Time Consuming, Lack of understanding of the context-specific words by authors
- Lack of sufficient information to distinguish between UA/UCs

  Found AG piping team working without dark hours risk assessment . Trialer driver moving the trialer without becon light.

- Insufficient Information, Referred to Corrective Actions

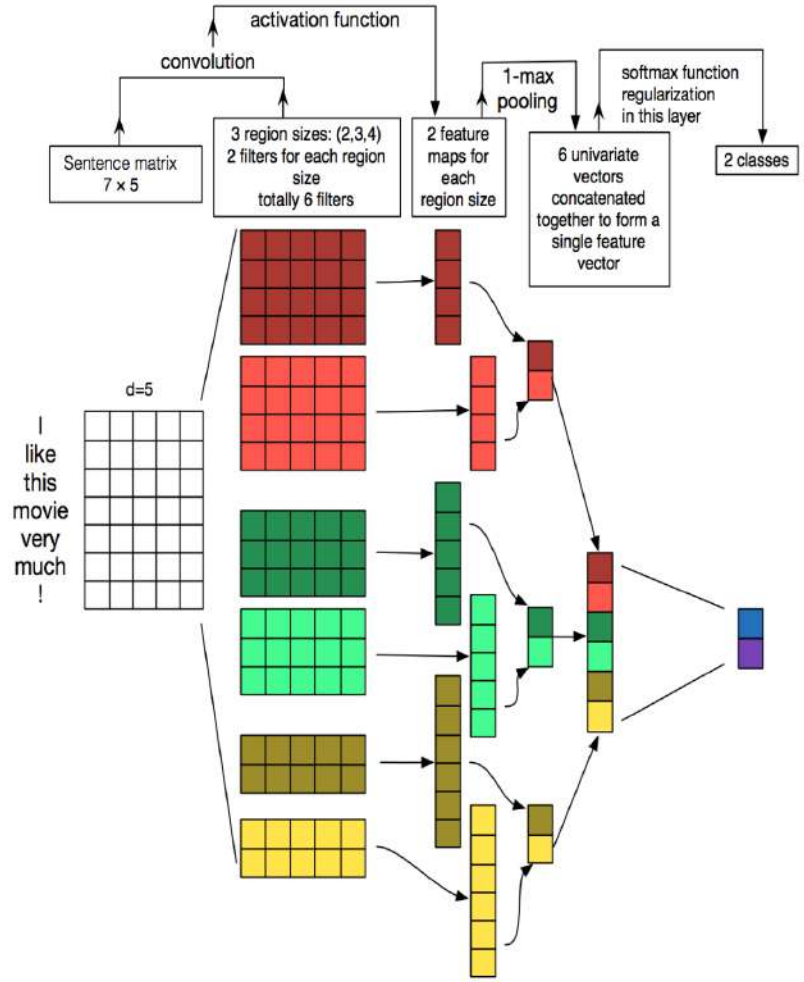
“Before sun rise wear white google afetr sun rise wear black google” - decided to remove

During the duct bank construction keep good arrangement for trafic for backfilling work.

CA - Good arrangements and wish them. Final Label - GO

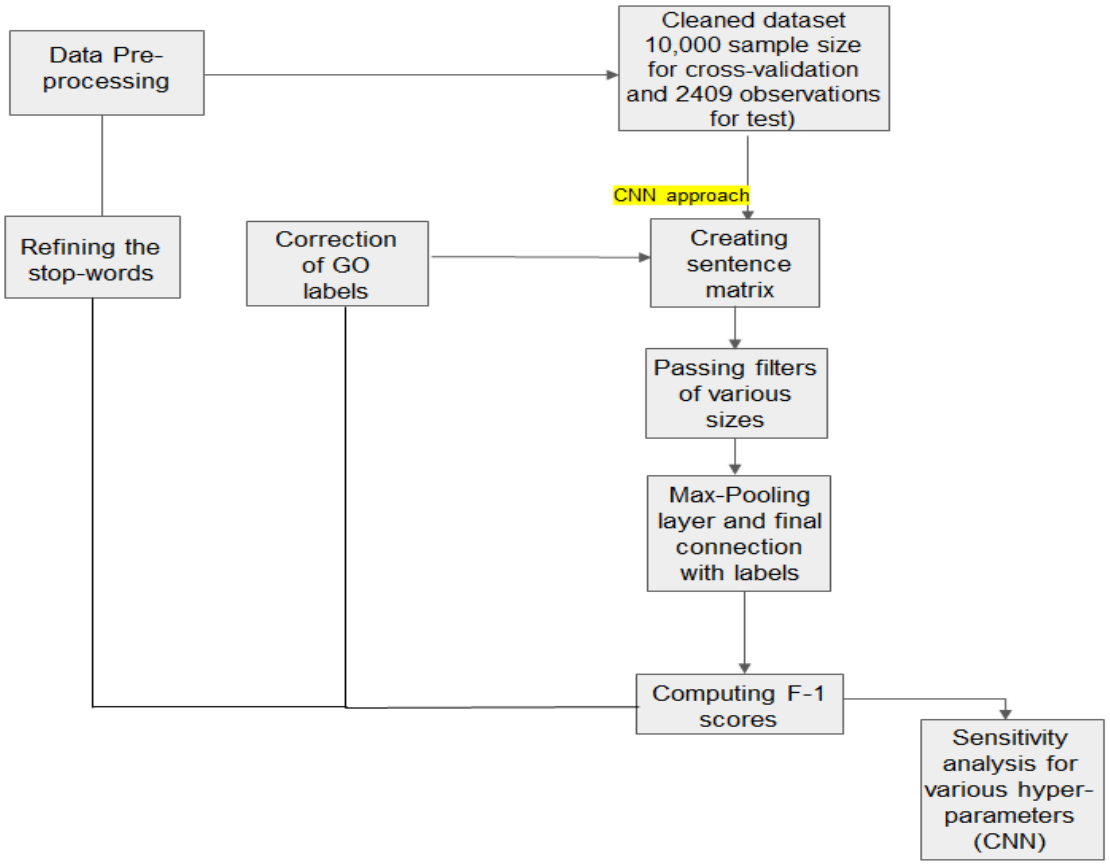
CNN approach

- Convolutional neural networks (CNNs) are the most widely used deep learning architectures in image processing and image recognition.
- The input layer for CNN for sentence classification is a sentence/document matrix. Each row of this input layer represents one token, that is each row is a vector that represents a word. Typically, we get these word vectors through word embeddings (low-dimensional representations) like word2vec.
- A weight matrix of nxn is then slid horizontally across the sentence by one step(also known as stride) capturing n words at a time. This weight matrix is called a filter.
- The central idea of pooling that we have to divide the output layers into subsections and calculate a value that best represents the output.
- The fully connected layer at the end receives the input from the previous pooling and convolutional layers, it then performs a classification task.



Convolutional Neural Network (CNN) architecture in NLP (Zhang & Wallace, 2015)

Methodology (CNN approach)



Results (CNN approach)

Table 8: Average F1 scores on 10,000 sample data

	GO	UA	UC	Average
Base Case	0.79	0.622	0.774	0.738
Stop Word Case	0.81	0.627	0.781	0.749
Mislabel Correction case	0.852	0.603	0.793	0.759

Table 9: Average F1 scores on 2409 test data

	GO	UA	UC	Average
Base Case	0.773	0.638	0.787	0.744
Stop Word Case	0.798	0.645	0.798	0.76
Mislabel Correction case	0.851	0.629	0.812	0.775

Non CNN vs CNN

- TF-IDF + SVM and other classifiers like RF lives up to their reputation, and reach very high performance everywhere. Interestingly, it even outperforms CNN in the mislabel correction case for UC label (see results section).

- However, our findings revealed that there is not a significant increase in the F-1 scores obtained from a more complicated and sophisticated neural network approach (CNN) compared to the traditional algorithms. (Baker et al., 2020) suggested that deep learning model could not stand out on their relatively small datasets (for deep learning standards). The sample size considered in that study was 90,000 (see table 2) and considering the dataset that we used, it is evident that the small data size is a reason was CNN to not perform significantly well from the other classifiers.

- There could be few broad reasons which could have been implemented to improve the F-1 scores:

  (i) Using algorithms to fine tune the hyper-parameters of the model.

  (ii) Use linguistic feature engineering to extract the connection in sentences and then use them as input to the neural network.

Conclusion

➔ Previous studies reported F1 score in the range of 0.55 -0.92 (Goy and Ubeynarayana 2017) for a structured data set.

➔ Preliminary results of the real data are comparatively acceptable though the data quality is poor and can be improved further.

➔ The improvement in the F-1 score using CNN is not as much as expected which could be attributed to the data quality, quantity and lack of proper hyper-parameter tuning .

WON the runners up for best technical paper at "The 6th PMI Research and Academic Virtual Conference 2021"

Project Management Institute.

HOSTED BY

IIT BOMBAY

THANK YOU