Centre for Research Training

Science Foundation Ireland sfi For what's next

# Offensive Language in Social Media

Duy Nguyen (DCU), Andrew Neary (NUIG), Alberto Castagna (TCD),
Andrea Rossi (UCC), Enda O'Shea (UL), Priya Rani (NUIG)

28th May 2021

HOST INSTITUTION

UCC University College Cork, Ireland Coláiste na hOllscoile Corcaigh

PARTNER INSTITUTIONS

University of Limerick Ollscoil Luimnigh

NUI Galway OÉ Gaillimh

DCU

Trinity College Dublin Coláiste na Tríonóide, Baile Átha Cliath The University of Dublin

CRT·AI SFI Centre for Research Training in Artificial Intelligence

# Problem Definition

**A. Offensive Language Identification**

- **OFF**ensive
- **NOT** offensive

**B. Categorization of offense type**

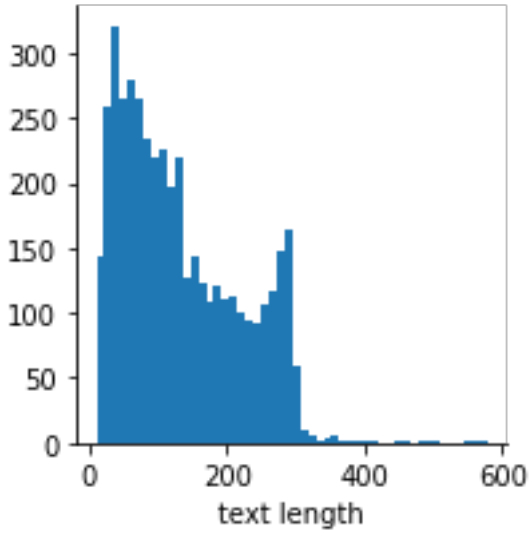
- **T**argeted **IN**sult
- **UNT**argeted

**OLIDv1.0 Dataset [9]**

| | id | tweet | subtask_a | subtask_b |
|---|---|---|---|---|
| 0 | 86426 | @USER She should ask a few native Americans wh... | OFF | UNT |
| 1 | 90194 | @USER @USER Go home you're drunk!!! @USER #MAG... | OFF | TIN |
| 2 | 16820 | Amazon is investigating Chinese employees who ... | NOT | NaN |
| 3 | 62688 | @USER Someone should'veTaken" this piece of sh... | OFF | UNT |
| 4 | 43605 | @USER @USER Obama wanted liberals & illega... | NOT | NaN |
| 5 | 97670 | @USER Liberals are all Kookoo !!! | OFF | TIN |
| 6 | 77444 | @USER @USER Oh noes! Tough shit. | OFF | UNT |
| 7 | 52415 | @USER was literally just talking about this lo... | OFF | TIN |
| 8 | 45157 | @USER Buy more icecream!!! | NOT | NaN |
| 9 | 13384 | @USER Canada doesn't need another CUCK! We alr... | OFF | TIN |

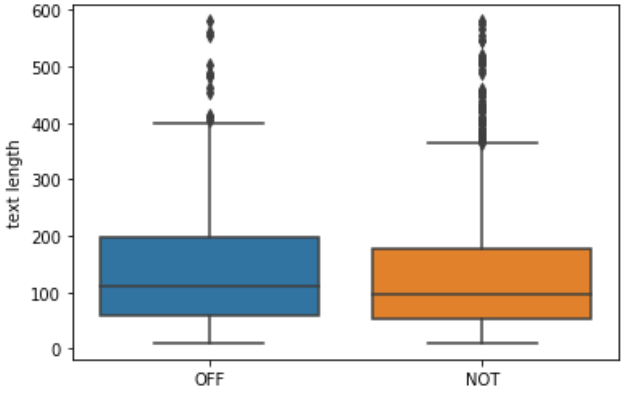
# Data Analysis

| Task A | Task B | Train | Test | Total |
|---|---|---|---|---|
| OFF | TIN | 3,876 | 213 | 4,089 |
| OFF | UNT | 524 | 27 | 551 |
| NOT | - | 8,840 | 620 | 9,460 |
| All | | 13,240 | 860 | 14,100 |

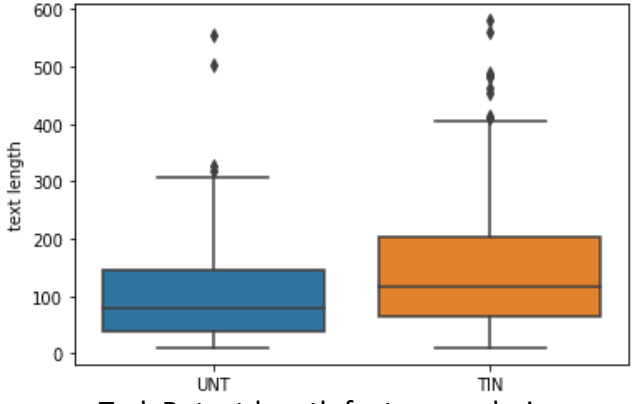




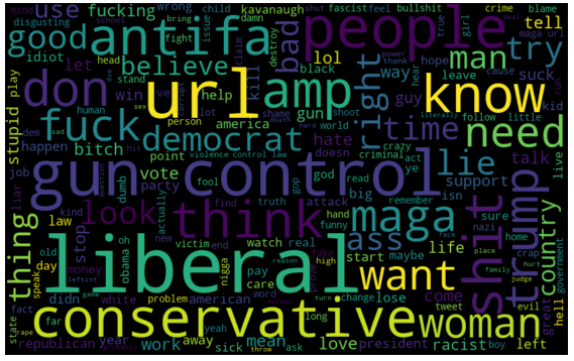
Most common words in **not offensive** tweets




Task A: text-length feature analysis




Task B: text-length feature analysis




Most common words in **offensive** tweets

# State of the art

- Deep learning techniques

  Bert [2,4], LSTM [6,7]

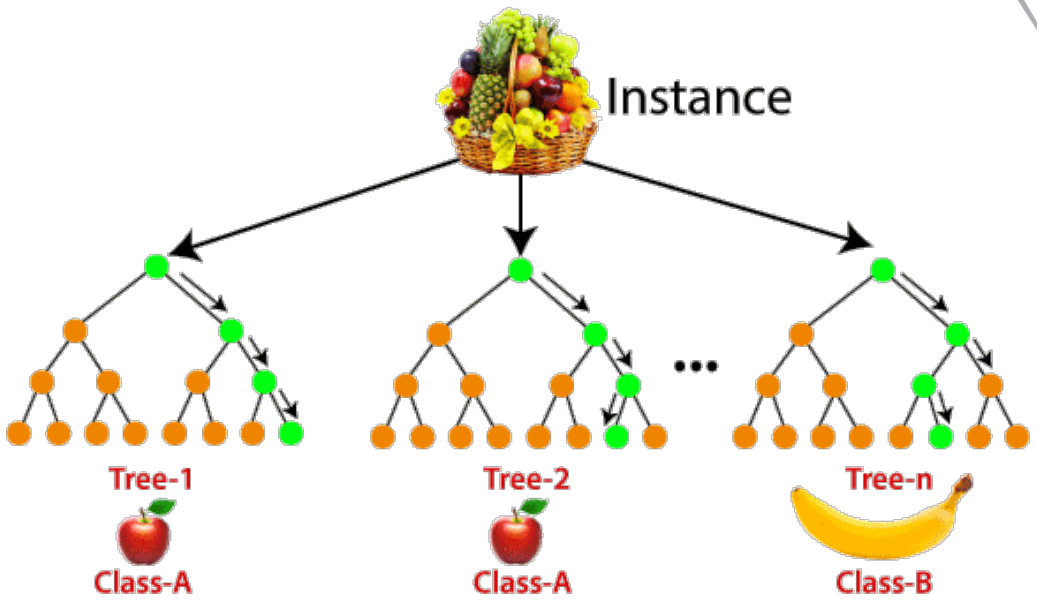
- Other ML techniques

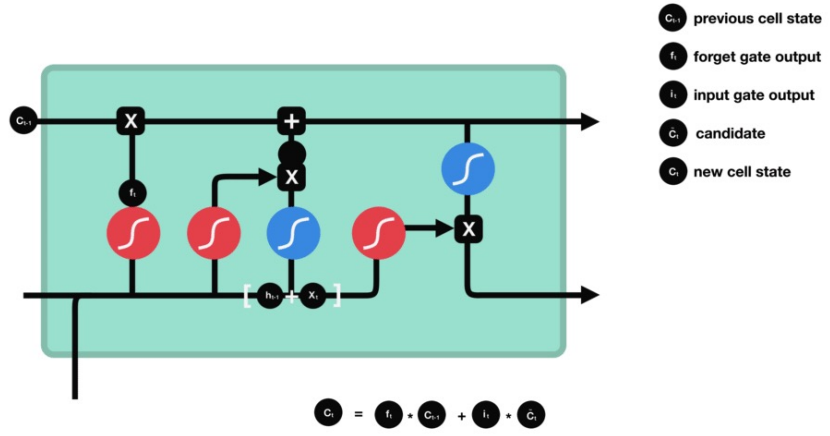
  Random forest, Logistic Regression [8]

- Sentiment analysis with user-related features (I.e., frequency of profanity in previous messages [5])

# Models

- **ML approaches**
  - Logistic Regression
  - Random Forest
  - XGB
- **DL approaches**
  - LSTM+CNN
  - BERT



# ML approaches: Pre-processing and Embeddings




- Stop word Removal
- Removal of extra spaces
- Removal of @name[mention]
- Removal of links[https://abc.com]
- Removal of punctuations and numbers
- Removal of emojis
- Tokenizing

- TF IDF
- Word2Vec

spaCy

https://spacy.io/

| | id | tweet | subtask_a | subtask_b | subtask_c | text length | final_tweet |
|---|---|---|---|---|---|---|---|
| 0 | 86426 | @USER She should ask a few native Americans wh... | OFF | UNT | NaN | 71 | ask native americans |
| 1 | 90194 | @USER @USER Go home you're drunk!!! @USER #MAG... | OFF | TIN | IND | 67 | home drunk maga trump url |
| 2 | 16820 | Amazon is investigating Chinese employees who ... | NOT | NaN | NaN | 182 | amazon investigate chinese employee sell inter... |
| 3 | 62688 | @USER Someone should'veTaken" this piece of sh... | OFF | UNT | NaN | 65 | vetaken piece shit volcano |
| 4 | 43605 | @USER @USER Obama wanted liberals & illega... | NOT | NaN | NaN | 72 | obama want liberal amp illegal red state |

# Logistic Regression, Random Forest and XGB Classifiers

Logistic Regression:

- Penalty: L2
- Solver: lbfgs

Random Forest:

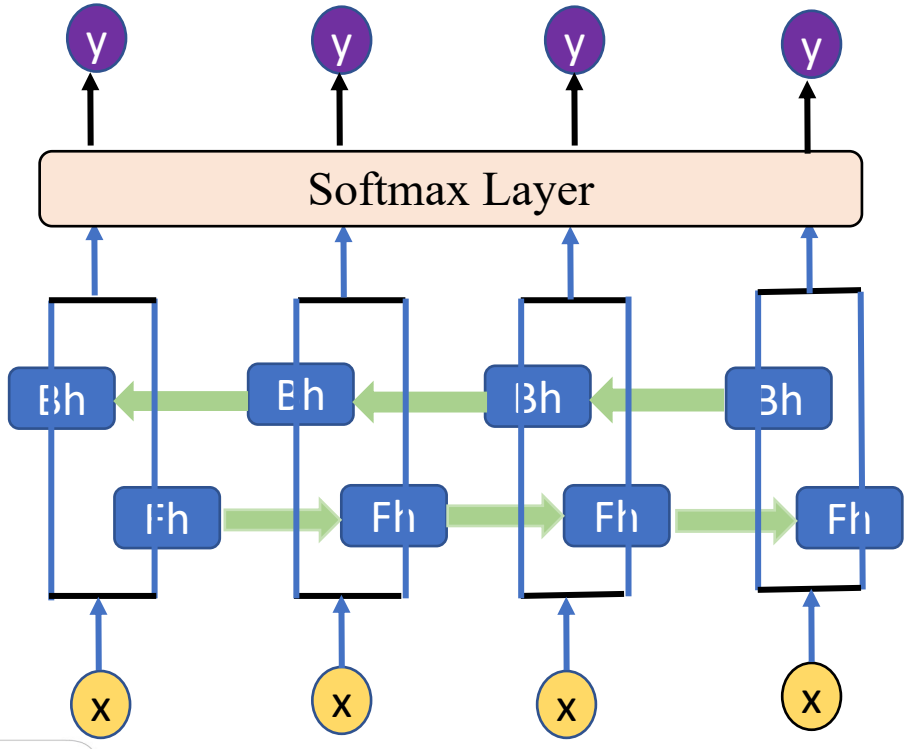
- Criterion: Gini impurity
- 100 estimators

XGBoost:

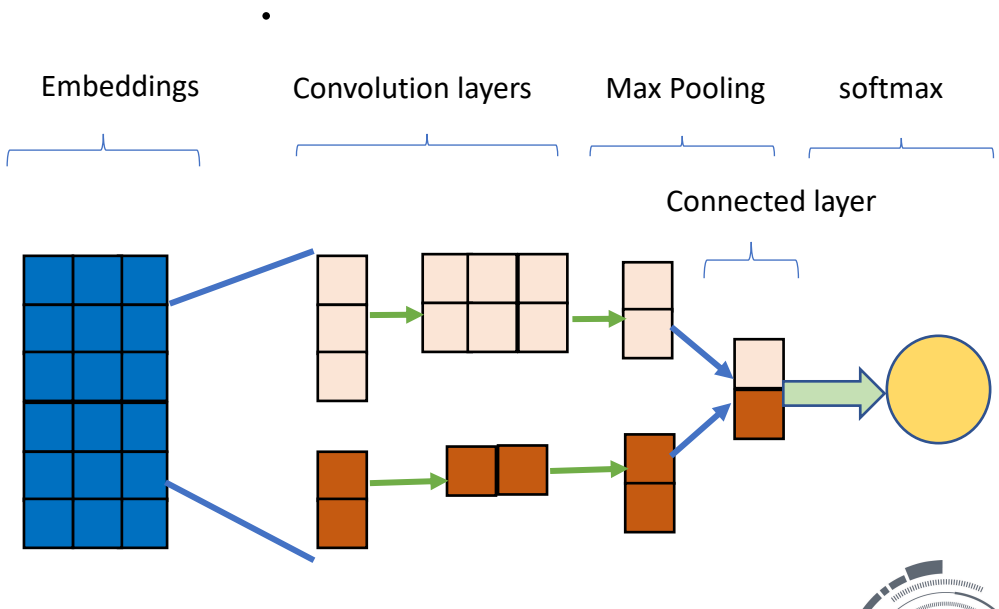
- Objective: Logistic regression for binary classification

# Neural Network Models

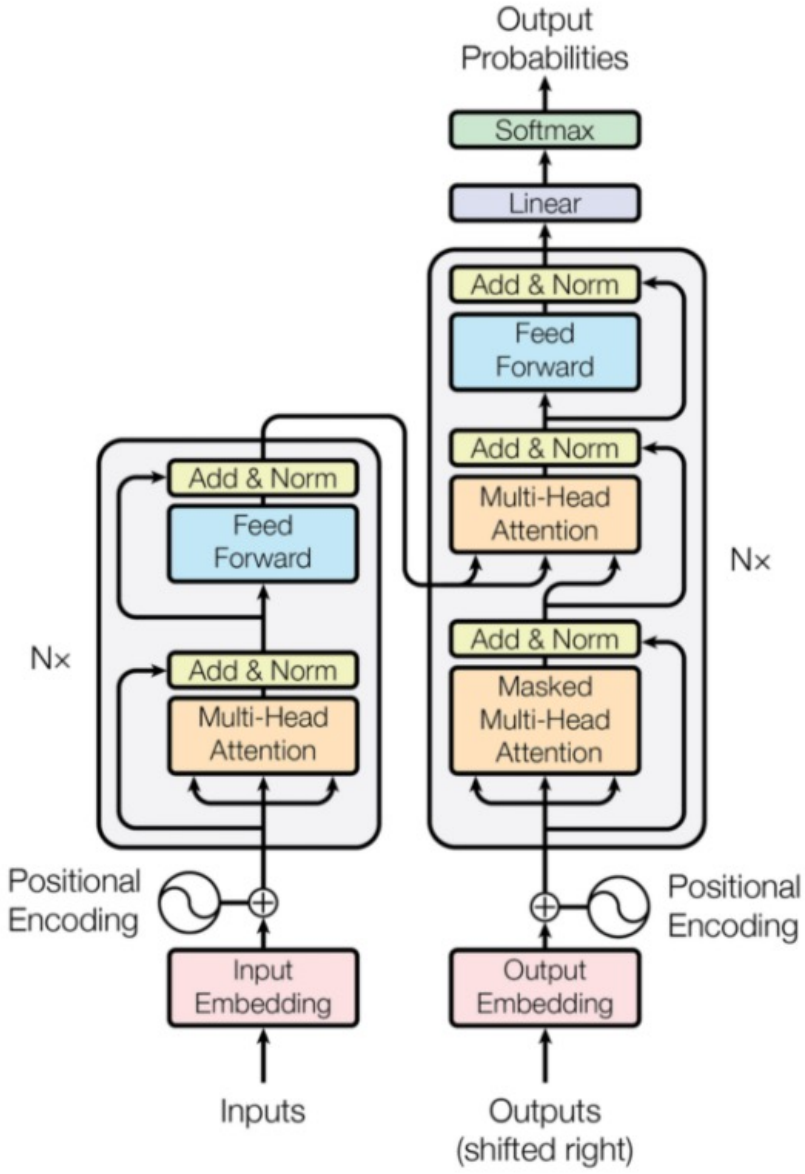
## Bi-LSTM Architecture




## CNN Architecture

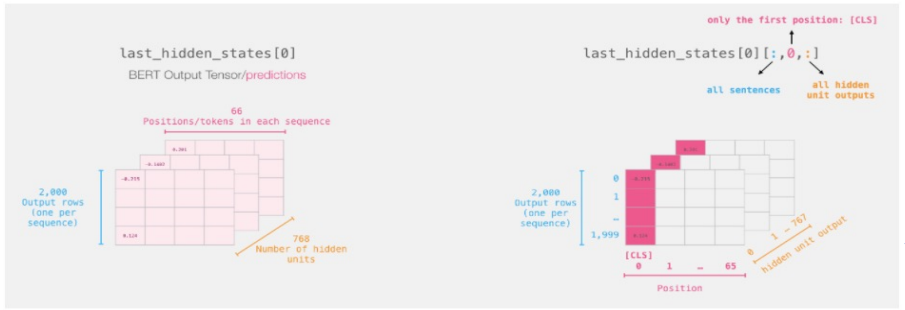


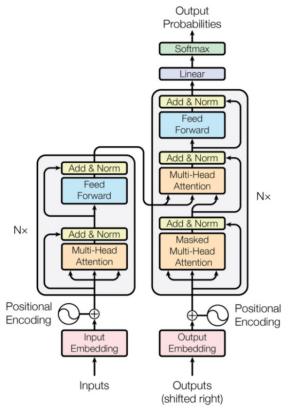
# Bidirectional Encoder Representations from Transformers (BERT)

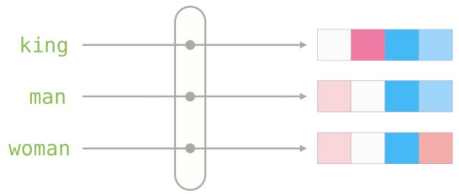


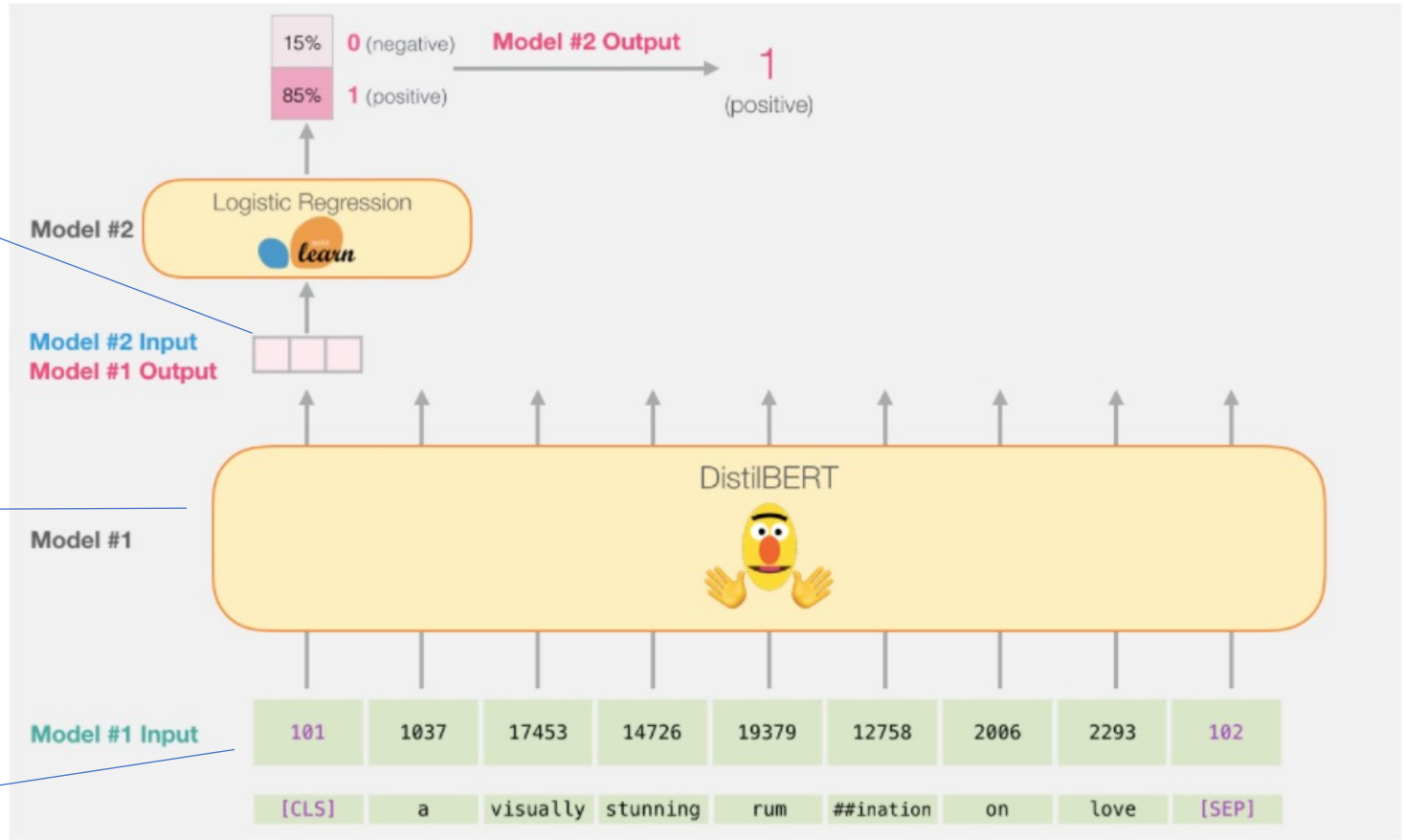

Reference [3] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L. and Polosukhin, I.

# distilBERT







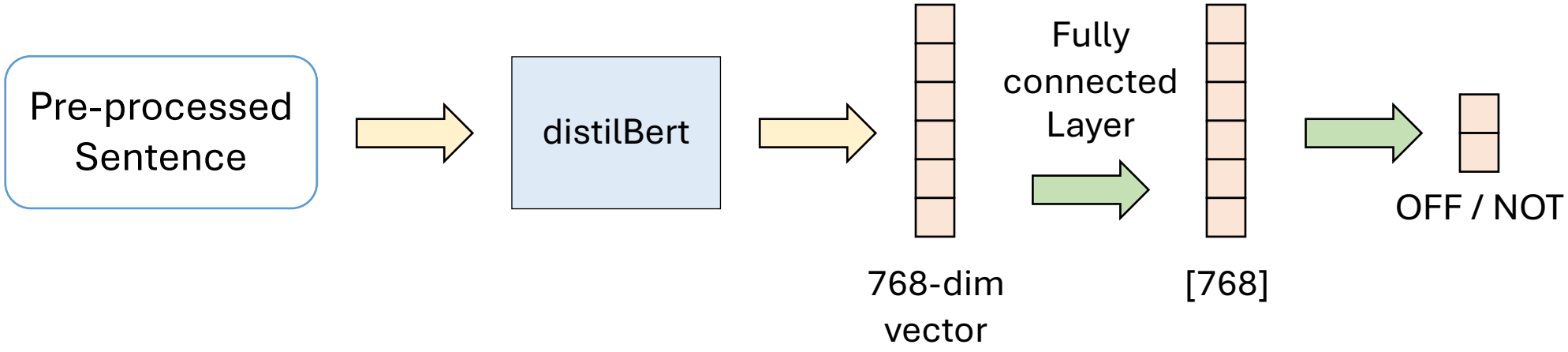
Reference [1] Alammar, J. [3] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L. and Polosukhin, I.

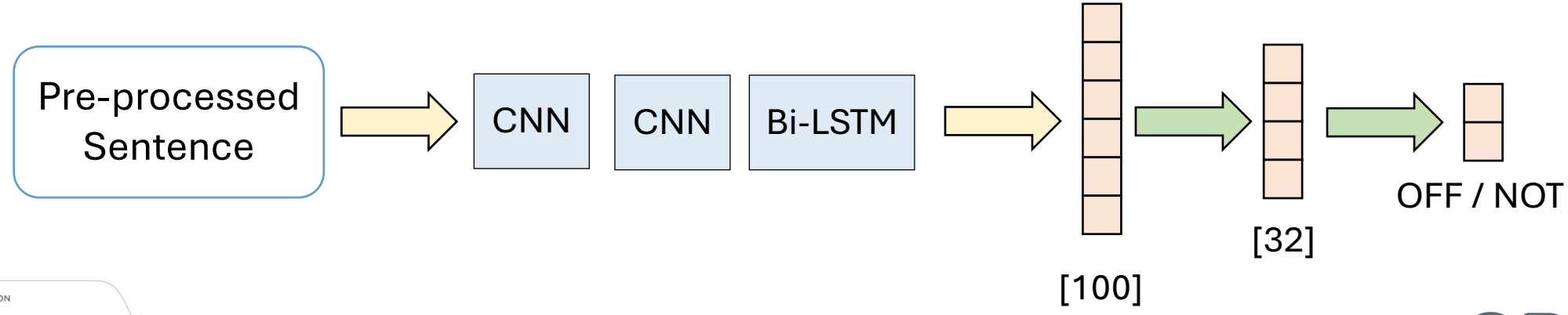
# Deep Learning Experiment

- (distilBERT) Pre-processing Data
  - Lowering
  - Transform emoji to text
  - (Remove emoji gave same result)

- (LSTM) Pre-processing Data
  - Lowering
  - Remove emoji to text
  - Removing URL

# Deep Learning Experiment

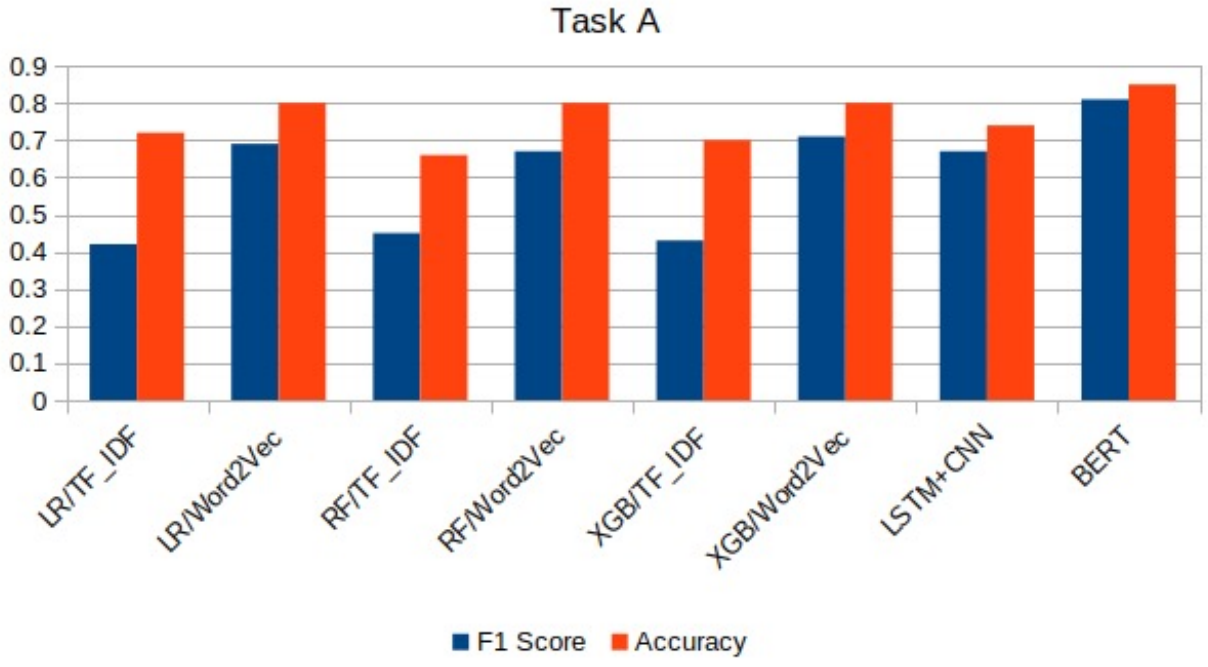


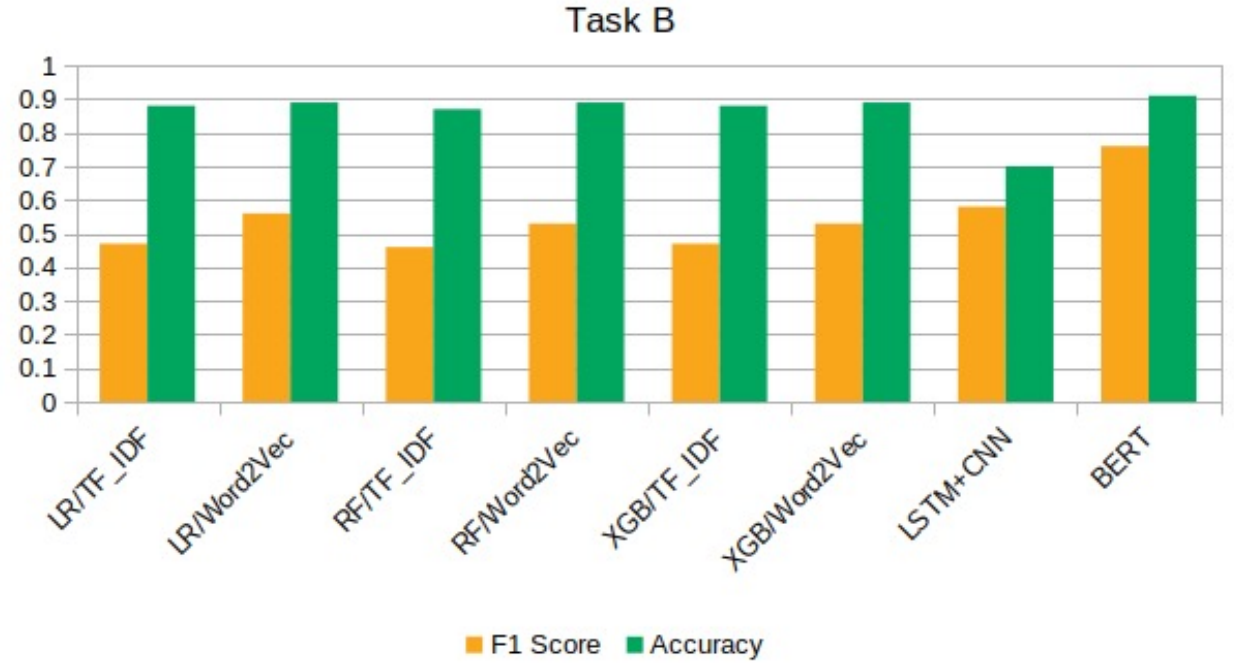



# Model Comparison

- Evaluation Metrics
  - Accuracy
  - F1
  - Precision
  - Recall

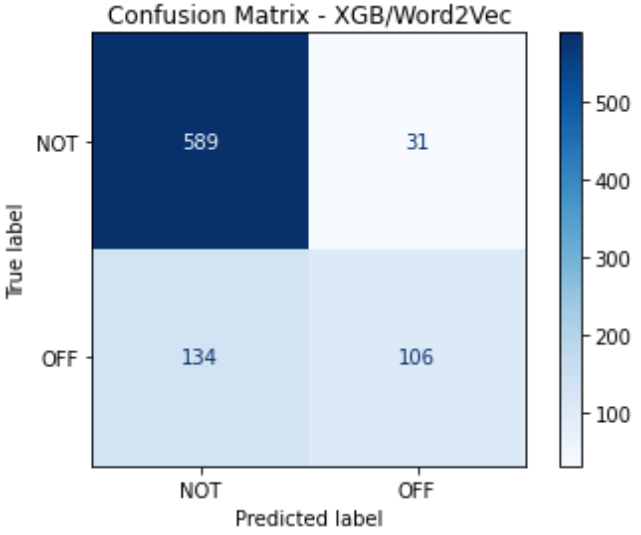
- Low F1 Score but high Accuracy
- DistilBERT gives the best F1 Score for both Task A & Task B
- Word2Vec works better than TF-IDF
- Not much difference in F1 score of Word2Vec and LSTM+CNN model



# Models Comparison – Subtask A

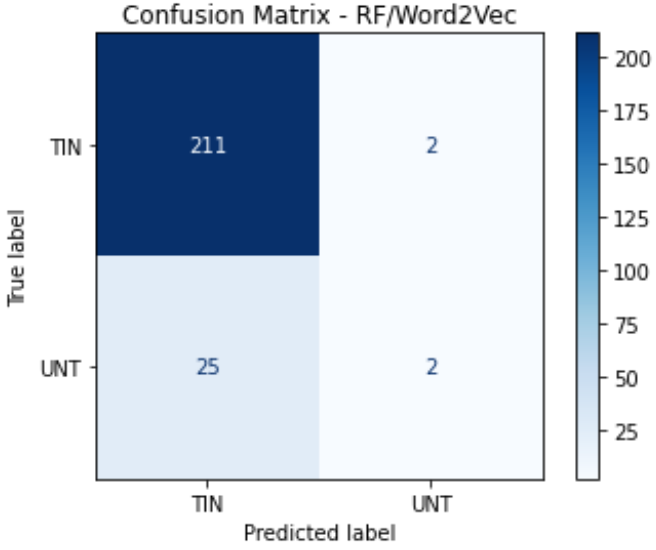
| Model | NOT | | | OFF | | | F1 Macro | Accuracy |
|---|---|---|---|---|---|---|---|---|
| | Precision | Recall | F1 | Precision | Recall | F1 | | |
| LR/TF_IDF | 0.72 | **0.99** | 0.83 | 0.25 | 0.008 | 0.02 | 0.425 | 0.72 |
| LR/Word2Vec | 0.80 | 0.97 | 0.88 | 0.83 | 0.36 | 0.50 | 0.69 | 0.80 |
| RF/TF_IDF | 0.7152 | 0.8870 | 0.7919 | 0.2307 | 0.0875 | 0.1269 | 0.4527 | 0.66395 |
| RF/Word2Vec | 0.7916 | 0.9806 | 0.8760 | **0.8695** | 0.3333 | 0.4819 | 0.6789 | 0.8 |
| XGB/TF_IDF | 0.7197 | 0.9774 | 0.8290 | 0.2222 | 0.0166 | 0.0310 | 0.43 | 0.7093 |
| XGB/Word2Vec | 0.8147 | 0.95 | 0.8771 | 0.7737 | 0.4416 | 0.5623 | 0.7197 | 0.8081 |
| LSTM+CNN | 0.75 | 0.91 | 0.82 | 0.68 | 0.41 | 0.51 | 0.67 | 0.74 |
| DistilBERT | **0.8752** | 0.9387 | **0.9058** | 0.8051 | **0.6542** | **0.7218** | 0.8138 | **0.8593** |
| Baseline | - | - | - | - | - | - | **0.829** | - |




| Label | Training | Test |
|---|---|---|
| OFF | 4400 | 240 |
| NOT | 8840 | 620 |

# Models Comparison – Subtask B

| Model | UNT | | | TIN | | | F1 Macro | Accuracy |
|---|---|---|---|---|---|---|---|---|
| | Precision | Recall | F1 | Precision | Recall | F1 | | |
| LR/TF_IDF | 0.0 | 0.0 | 0.0 | 0.8875 | **1.0** | 0.9404 | 0.4792 | 0.8875 |
| LR/Word2Vec | **0.75** | 0.1111 | 0.1935 | 0.90 | 0.9953 | 0.9443 | 0.5689 | 0.8958 |
| RF/TF_IDF | 0.0 | 0.0 | 0.0 | 0.8861 | 0.9859 | 0.9333 | 0.4667 | 0.875 |
| RF/Word2Vec | 0.6666 | 0.0741 | 0.1333 | 0.8945 | 0.9953 | 0.9422 | 0.5378 | 0.8917 |
| XGB/TF_IDF | 0.0 | 0.0 | 0.0 | 0.8875 | 1.0 | 0.9404 | 0.4702 | 0.8875 |
| XGB/Word2Vec | 0.6666 | 0.0741 | 0.1333 | 0.8945 | 0.9953 | 0.9422 | 0.5378 | 0.8917 |
| LSTM+CNN | 0.39 | 0.26 | 0.31 | 0.77 | 0.86 | 0.81 | 0.58 | 0.70 |
| DistilBERT | 0.625 | **0.5555** | **0.5882** | **0.9444** | 0.9577 | **0.951** | **0.7696** | **0.9125** |
| Baseline | - | - | - | - | - | - | 0.755 | - |




| Label | Training | Test |
|---|---|---|
| TIN | 3876 | 213 |
| UNT | 524 | 27 |

# Misclassified tweets

ID **60133** – Label: OFF, Prediction: NOT


*"¡No pasarán!" was used by British anti-fascists during the October 1936 Battle of Cable Street, and is still used in this context in some political circles. It was often accompanied by the words nosotros pasaremos (we will pass) to indicate that communists rather than fascists will be the ones to seize state power.* - Wikipedia

- Out of vocabulary
- Spanish language

ID **54053** – Label: NOT, Prediction: OFF


- The word itself it's considered offensive

# Conclusions

- Pre-trained models are extremely valuable in NLP
- Challenges and future work:
  - Imbalanced classes (better dataset)
  - Hyperparameters tuning
  - Additional pre-processing steps
  - Ambiguous data – is binary classification reasonable?

# References


1. Alammar, J., 2021. *A Visual Guide to Using BERT for the First Time*. [online] Jalammar.github.io. Available at: <https://jalammar.github.io/a-visual-guide-to-using-bert-for-the-first-time/> [Accessed 27 May 2021].

2. Horev, R., 2021. *BERT Explained: State of the art language model for NLP*. [online] Medium. Available at: <https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270> [Accessed 27 May 2021].

3. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L. and Polosukhin, I., 2021. *Attention Is All You Need*. [online] arXiv.org. Available at: <https://arxiv.org/abs/1706.03762> [Accessed 27 May 2021].

4. Devlin, J., Chang, M., Lee, K. and Toutanova, K., 2021. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. [online] arXiv.org. Available at: <https://arxiv.org/abs/1810.04805> [Accessed 27 May 2021].

5. Dadvar, Maral & Trieschnigg, Dolf & Ordelman, Roeland & de Jong, Franciska. (2013). Improving Cyberbullying Detection with User Context. In Proceedings of 35th European Conference on IR Research, ECIR 2013, Advances in Information Retrieval. pp 693-696. 10.1007/978-3-642-36973-5_62.

6. M. Susanty, Sahrul, A. F. Rahman, M. D. Normansyah and A. Irawan, "Offensive Language Detection using Artificial Neural Network," 2019 International Conference of Artificial Intelligence and Information Technology (ICAIIT), 2019, pp. 350-353, doi: 10.1109/ICAIIT.2019.8834452.

7. Bisht, Akanksha, Annapurna Singh, H. S. Bhadauria, and Jitendra Virmani. "Detection of hate speech and offensive language in twitter data using lstm model." In *Recent Trends in Image and Signal Processing in Computer Vision*, pp. 243-264. Springer, Singapore, 2020.

8. Pedersen, Ted. "Duluth at SemEval-2020 Task 12: Offensive Tweet Identification in English with Logistic Regression." *arXiv preprint arXiv:2007.12946* (2020).

9. https://sites.google.com/site/offensevalsharedtask/olid [Accessed 25 May 2021].

# Acknowledgements

Thank you for your attention!
Any questions?


https://github.com/andreareds/NLPWeek_Offensive_language

Science Foundation Ireland Grant No- 18/CRT/6223