CORK INSTITUTE OF TECHNOLOGY
INSTITIÚID TEICNEOLAÍOCHTA CHORCAÍ

http://www.cit.ie

Offensive Language Detection in Social Media Posts.

Sidharth Mehra

27th May 2020

Supervisor

Dr. Mohammed Hasanuzzaman

Context and Objectives of the Research Topic

- Predicting whether a social media post is offensive or not with the help of NLP, Machine Learning and Deep Learning.
- Handling Imbalance in the twitter dataset is one of the main objective.
- Comparative Analysis of various Feature Extraction techniques on the OLID twitter dataset.
- Exploration of range of several Machine Learning Algorithms for detecting the nature of the social media post.
- Application of Deep Learning Algorithms to explore the performance on the newly created dataset.

Motivation to take this Research Problem ?

- Social media sites like Facebook and Twitter have become much popular due to the seamless waive of Internet.
- Posting Offensive Content on these platforms has also increased leading to the misuse of the freedom of speech.
- Offensive Posts and Comments on these sites is causing a huge social unease and negatively impacting the victimized individual or any community.
- Because of the abusive content on these platforms, there are several extreme cases of mental depression, distress and suicidal attempts.
- Detecting the nature of the posted content on these platforms will automate and accelerate the task of normalizing these posts.
- Efficient NLP and AI systems have the capability to detect these type of offensive content so as to take the relevant action as quick as possible and minimize this unease and save lives.

Approach Undertaken to solve the Research Problem

Machine Learning Lifecycle for Text Classification

- **Exploratory Data Analysis** (exploring the dataset to get better insights from the various statistics)
- **Text Pre-processing** (cleaning the twitter data for model building)
- **Feature Extraction** (producing the numerical representation of the text data)
- **Handling Imbalance** (making the training data distribution equal)
- **Model Building** (various ML and DL models such SVM, Logistic Regression, Random Forest, ANN, CNN, LSTM's)
- **Model Evaluation** (evaluating the model detection performance on the test set using useful metrics such as F1-score and confusion matrix)

Background and Literature Review

- Explored 15 academic research papers revolving around the project topic in terms of motivation and challenges, datasets involved, existing methodology, evaluation and result analysis.
- All the previous studies focused on either of 5 categories revolving around the problem statement which were Cyber-bullying, Hate-speech, Aggression, Toxic Comment Classification and Overall offensive content.
- Only one recent paper in 2019 released a dataset named OLID (Offensive Language Identification Dataset) which considered the problem as the whole by taking all the categories of the offensive content into consideration.

Background and Literature Review Continue…

- OLID 2019 twitter dataset aims to capture the differences and similarities between the pre-existing datasets revolving around the mentioned 5 categories that fall into the topic of detecting the offensive content on the social media.
- Moreover this dataset enables the identification of the type and the target of the detected offensive tweets for further analysis and actions.
- This paper gave us the baseline results on the novel dataset and therefore opened up the new research directions to meaningfully contribute in the field.

Implementation and Methodology

- As the dataset is new and not much work has been done on this, we have explored various techniques in the each step of our project methodology.
- **Exhaustive Text Pre-processing** – Carried out using NLTK Library
- Below are the steps that are followed in cleaning our messy text data

Stemming

Lower Case Conversion

Tokenisation

Lemmatization

URL mentions Removal

Stop-words Removal

Punctuation Removal

Contraction Expansion

Special Characters Removal

Number Handling

White Space Removal

Emojis Handling

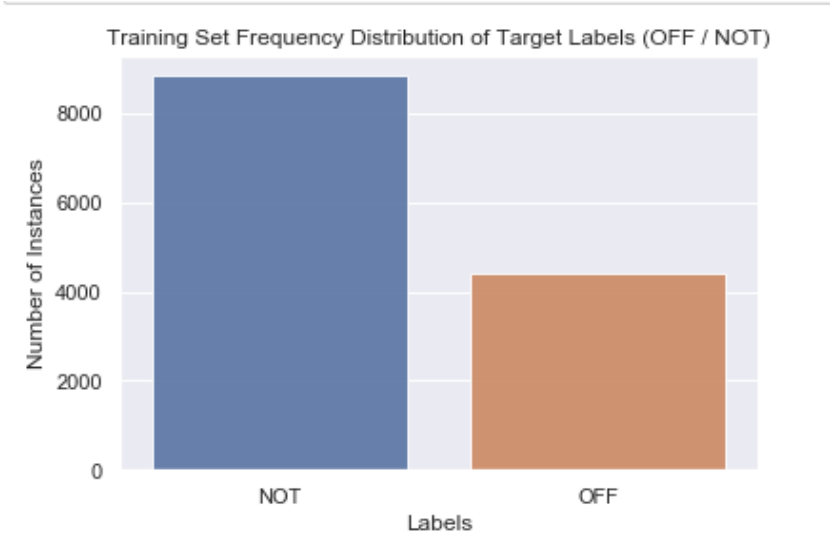
Implementation and Methodology…

- Next we proceeded with experimenting with various **Feature Extraction Mechanisms.**
- Below are some techniques to extract the important features which we explored on our benchmark Dataset OLID.

Statistical Models for Feature Extraction	Word Embeddings for Feature Extraction
1) Bag of Words	1) Word2Vec
2) Tf-Idf	2) GloVe
3) BOW Combinations of Uni-gram, Bi-gram and Tri-gram. 4) Tf-Idf Combinations of Uni-gram, Bi-gram and Tri-gram.	3) BERT

Implementation and Methodology…

- After the results from Exploratory Data Analysis, we found that our training dataset was imbalanced.
- 33% of the tweets were Offensive and 67% of the tweets were Non-Offensive.
- Below figure shows the imbalance problem in the OLID dataset which needs to handled otherwise the results will be biased towards majority class only.



Implementation and Methodology…

- In the next step we proceeded with handling imbalance in our training dataset.
- We explored various over-sampling and under-sampling techniques to mitigate this problem of imbalance.

Over-sampling Techniques	Under-sampling Techniques
1) SMOTE	1) Tomek Links
2) Random Over-sampling	2) Random Under-sampling
3) ADASYN	3) Cluster Centroids

- Random Under-sampling performed best with all the feature extraction and model building algorithms.

Implementation and Methodology…

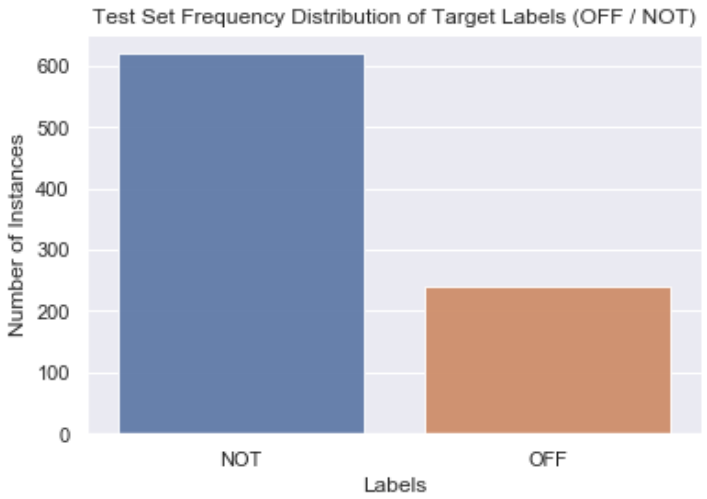
- Further in the process flow of detecting the nature of the social media posts we have build a range of Machine learning and Deep learning Models.

Machine Learning Models	Deep Learning Models
1) Support Vector Machines 2) Logistic Regression 3) Naïve Bayes 4) Decision Tree 5) Random Forest 6) Gradient Boosting 7) Ada Boost	1) Artificial Neural Network 2) CNN + LSTM 3) CNN + Bi-LSTM 4) BERT

- We evaluated the performance of previously mentioned feature extraction techniques with each of these Machine learning and Deep learning models.

Implementation and Methodology…

- **Model Evaluation:** This is the process of measuring the performance of our machine learning or deep learning models.
- Our test is imbalanced, so the normal accuracy cannot be chosen as the appropriate metric.



- Therefore we choose Confusion Matrix and Macro F1-score to evaluate the goodness of our models.

Comparative Analysis and Results

- After evaluating the different feature extraction techniques along with the various ML and DL algorithms, we come up with the comparative analysis of the top performing models.
- We report the best Feature Extraction Mechanisms along with the best Machine Learning and Deep Learning Models on our novel OLID 2019 dataset.
- Next slide present the final results evaluated on the OLID test set.
- We used the insights from confusion matrix and Macro F1 score to come up with the best model for our problem domain.

Comparative Analysis and Results…

Top Performing Models	Macro F1- Score	Majority Class Accuracy	Minority Class Accuracy
1) BOW (Unigram) + SVM	0.70	0.77	0.67
2) BOW (Unigram + Bigram + Trigram) + Decision Tree	0.68	0.73	0.67
3) Tf-Idf (Unigram + Bigram + Trigram) + Logistic Regression	0.70	0.77	0.65
4) Tf-idf (Unigram) + Naïve Bayes	0.68	0.70	0.72
5) Word2Vec + Logistic Regression	0.74	0.72	0.74
6) GloVe + SVM	0.75	0.80	0.71
7) BOW + CNN-BiLSTM	0.77	0.89	0.64
8) BERT	0.82	0.91	0.70

Comparison with the Baseline Results

	NOT			OFF			Weighted Average			
Model	P	R	F1	P	R	F1	P	R	F1	F1 Macro
SVM	0.80	0.92	0.86	0.66	0.43	0.52	0.76	0.78	0.76	0.69
BiLSTM	0.83	0.95	0.89	0.81	0.48	0.60	0.82	0.82	0.81	0.75
CNN	0.87	0.93	0.90	0.78	0.63	0.70	0.82	0.82	0.81	**0.80**
All NOT	-	0.00	0.00	0.72	1.00	0.84	0.52	0.72	0.	0.42
All OFF	0.28	1.00	0.44	-	0.00	0.00	0.08	0.28	0.12	0.22

Table 4: Results for offensive language detection (Level A). We report Precision (P), Recall (R), and F1 for each model/baseline on all classes (NOT, OFF), and weighted averages. Macro-F1 is also listed (best in bold).

- Our results with GloVe + SVM achieved the F1 Macro of 0.75 as compared with the baseline score of 0.69.
- Application of the state of the art technique BERT on the dataset resulted in the overall F1 Macro of 0.82.

Conclusion

- Extensive literature review helped us to explore the recent OLID 2019 twitter dataset for offensive language detection.
- After trying out various techniques for handling the imbalance in the training set, random under-sampling proved to be the best one.
- Exhaustive Comparative Analysis of various Feature Extraction mechanisms, ML and DL algorithms on the dataset.
- BERT came out to be the best model achieving the F1 Macro of 0.82 which beats the performance of the baseline work.

Future Work

- Focus on sub-task B (type of offensive post) and sub-task C (target of the targeted offensive post)
- Capture syntactic and semantic features along with their combination and other pre-trained features.
- Under-sampling leads to the loss of valuable information, therefore explore the suitable method to tackle the imbalance issue in our dataset which preserve this information.
- Develop the ensemble of ML and DL models.
- Increasing the complexity of the Deep Learning models along with hyper parameter optimization.

CORK INSTITUTE OF TECHNOLOGY
INSTITIÚID TEICNEOLAÍOCHTA CHORCAÍ

http://www.cit.ie

Thank you

Sidharth Mehra

27th May 2020