CIL: Text Sentiment Classification
Group: Shazam!

Noman Ahmed Sheikh, Rishabh Singh, Samriddhi Jain, Kumar Mohit
Dept. of Computer Science
ETH Zurich

***Abstract*—Twitter sentiment analysis continues to be an exigent exercise owing to the diversity and subjectivity involved. Existing state-of-the-art methods are derived from transformers, however they fail to reach a consensus over sentiment polarity of many tweets containing polysemous words. We propose an ensemble classifier based on BERT and RoBERTa and argue for its consistency and robustness across multiple examples. The incisive analysis of the results obtained echo the efficacy of our proposal.**

I. Introduction

Social media applications have dramatically burgeoned over the past decade. These platforms have changed internet from a static repository of information to a dynamic forum of continuously updating data. Such data is used to analyse human behavior such as customer review, public opinion etc. and draw apt conclusions. Sentiment analysis, which is the computational study of people's opinions, appraisals, attitudes, and emotions toward entities, individuals, issues, events, topics and their attributes [1], lies at the core of these works. Twitter is the most popular micro-blogging platforms, with roughly 500 million messages per day [2]. The fact that access and download of published posts is provided with ease, Twitter is considered one of the largest datasets of user generated content. However, analysing *tweets* comes with some inherent challenges such as text length (140 characters max), grammar, data sparsity (great percentage of tweets' terms occur fewer than 10 times [3]), presence of negation, stop words, tokenization, multilingual content, multimodal content, breadth of domain(people tweet about anything and everything [4]) etc. There have been a plethora of attempts at addressing these challenges- semantic smoothing to reduce the sparseness [3]), contextual negation [5], common list of stopwords across multiple datasets [6], Twitter-specific tokenizer [7], language-independent classifier [8] etc. However, inferring accurate sentiment from tweets remains unfathomable.

A. Related Work

[2] organises the existing literature on twitter sentiment analysis into four broad categories- Machine Learning, Lexicon-based, Hybrid and Graph-based methods. The recent advent of deep learning methods has seen a shift from hand-crafted features to learnt representations such as Word2Vec [9] and Glove [10]. Recursive Neural Networks is a class of architecture that can learn a directed acyclic graph structured input (e.g., a tree structure) [11]. They have been widely used for sentence level [12], phrase-level [11] and aspect-level [13] sentiment analysis. A myriad of state-of-the-art methods have been using Recurrent Neural Networks such as target-dependent LSTM [14], GRU to model syntax and semantics [15], hierarchical bidirectional LSTM [16] etc. Attention has been another success story in many natural language processing tasks, and twitter sentiment analysis is not untouched either. Many studies have focused on improving the basic attention RNN models [10] while an equally exciting number of works have concentrated on interactive attention-based RNN models [17]. Convolutional Neural Networks are able to extract both local and global representations from a text, precisely why numerous works leverage CNN for incorporating aspect information [18], positional relevance [19] etc. Finally, memory networks along with attention mechanism have obtained great success over the years in twitter sentiment analysis [20], [21]. The major breakthrough in language modelling came with the introduction of BERT [22] which used masked language models and next sentence prediction techniques to learn a sentence encoding, which can be later fine-tuned to specific tasks. Soon came the tide of different transformers based models, notable among these being, RoBERTa [23], XL-Net [24] and XLM RoBERTa [25]. RoBERTa introduced dynamic masking and improved training techniques over BERT, whereas XLNet introduced permutation based sequential token learning using TransformerXL [26]. XLM RoBERTa is a multi lingual model, based on RoBERTa, which is designed to understand the language directly from input. All these models showed significant improvement over BERT and are considered benchmarks now in different domains.

We observe that model-based methods are accurate and robust, given adequate data is available. Moreover, deep learning outperforms conventional learning-based methods for twitter sentiment analysis. Further, ensemble methods tend to achieve higher performance than individual classifiers. We thus build upon these observations and propose an ensemble model of BERT and RoBERTa based classifiers for predicting the sentiment of tweets. We discuss our proposed methodology in the next section followed by a

detailed overview of the experimental setup. We conclude with a rigorous analysis of the results obtained and argue analytically the effectiveness of our proposed method.

II. Methodology

A. Pre-processing

In the pre-processing stage, we clip and pad all the tweet texts to a common character length of 140. We next use wordsegment library [27] to break the words and hashtags into further smaller words. Next, we tokenize the text by following WordPiece Tokenization as used in BERT. It greedily breaks the text into into words present in the vocabulary.

1) Word Segmentation: We found this stage to give a significant boost in performance on twitter data. Twitter usually consists of large number of hashtags and a majority of these hashtags are phrases containing two or more words concatenated together such as "#GetWellSoon", "#SocialDistancing" and so on. In order to enrich classifier with more useful information, we decided to segment such phrases into individual words rather than passing the entire phrase (as a word) to the tokenizer. We present an example where the effect can be seen more clearly. In the table mentioned below, Tweet 1 is the actual tweet which is classified incorrectly carries most of the sentiment information in the hashtag and the model is unable to extract it. Whereas in Tweet 2, segmenting hashtag into individual words leads to correct classification.

Id	Content	Prediction
Tweet 1	#MakeAmericaGreatAgain is what he said	Negative
Tweet 2	make america great again is what he said	Positive

B. Proposed model

Our best performing models are mostly based on transformers. We experiment with multiple transformer implementations such as BERT, RoBERTa and XLMRoBERTa. We experiment with taking multiple ensembles of these model variants and found that a combination of RoBERTa and BERT gives best results on kaggle as well as on our validation split. There are multiple ways to take ensemble of trained models such as (1) Taking logits of model output and converting them to prob. distribution using a softmax followed by averaging over all the models (2) Taking majority of predictions of the all models. (3) One can also adopt more complex approach such as extracting features by removing last layer of all the models and passing the concatenated features to an additional network that summarizes all the models. We replaced this additional network in multiple ways such as (1) adding a Bidirectional LSTM followed by a linear layer for classification. (2) Taking the output corresponding to CLS token and add 3 linear layers. While these approaches seems promising we stick to first approach focusing on improving individual models.

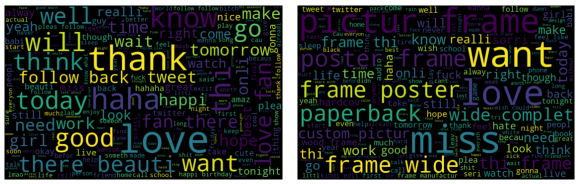


Figure 1: Word clouds of positive (left) and negative (right) sentiments in training data

III. Experiments & Results

A. Dataset

The full data for twitter sentiment analysis on Kaggle consist of a training set having a total of 2.5M tweets. Each tweet is annotated as either positive or negative, depending on the sentiment it conveys. We split this set into 99:1 ratio in order to create a validation set and used it to tune the hyper-parameters of our model. Rudimentary analysis of the dataset reveals the most frequent words (Fig 1) used for each class . This helps in understanding the dataset better.

B. Baselines

We compare our proposed method with a number of standard models and methods.

- Bag of Words features for the two classes and a logistic regression as the classifier.
- Term Frequency- Inverse Document Frequency features with logistic regression classifier.
- Glove [28] 200-dimensional word embedding with a logistic regression classifier.
- Hierarchical Attention Network [29] uses bidirectional GRU-based encoders along with contextual attention mechanism at both sentence and word levels.

Table I: Accuracy comparison with baseline models

	Models	Accuracy
Baselines	BoW	0.7017
	TF-IDF	0.7084
	Glove	0.7714
	HAN	0.8301
CNN models	Spacy-CNN	0.8480
Attention based models	BERT	0.8940
	RoBERTa	0.8994
	XLNet	0.8784
	XLM RoBERTa	0.8852
Experiments	RoBERTa + 3 FC layers	0.8892
	RoBERTa + bidirectional LSTM + 1 FC layer	0.8608
	RoBERTa + 3 Conv layers + 1 FC layer	0.8620
	Final proposed ensemble	**0.9086**

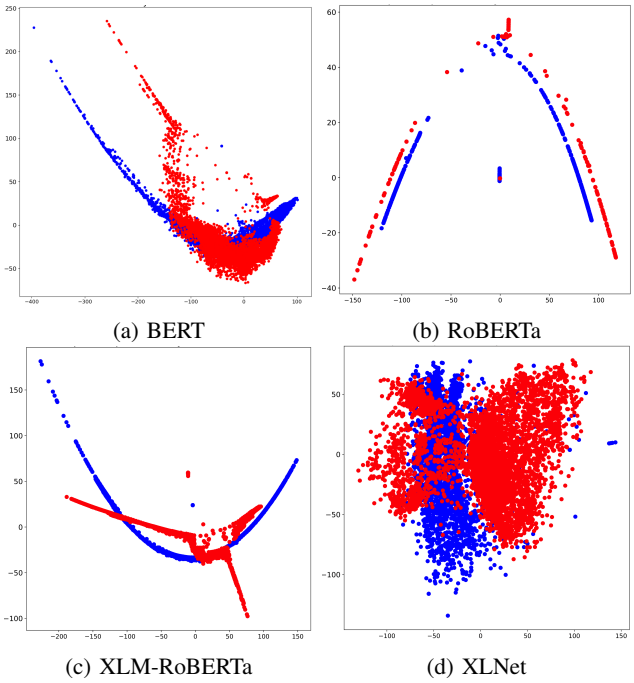
C. Results

We tabulate the accuracy obtained for different methods in Table I. We commenced with creating word2vec representation with CNN-based model using the library Spacy [30]. We observed a slight improvement of 2% over HAN. We then focused on attention-based models. We evaluated pre-trained versions of transformer-based methods- BERT, RoBERTa, XLNet and XLM-RoBERTa. It was observed that BERT and RoBERTa performed consistently well among the rest of the methods, with RoBERTa achieving approximately 28% over BoW and 8% over HAN. These results motivated us to understand why certain models perform better. Hence, we visualise the features extracted before the classification layer of each model. We use isomap, which is a nonlinear dimensionality reduction method, for visualizing the 768 dimensional features. In Figure 2 we observe that data from the two different categories are highly separable in case of RoBERTa as compared to XLNet. The feature separation between classes for BERT and XLM-RoBERTa is almost similar. These findings galvanised us to try configuring an ensemble model of RoBERTa and BERT. Since the feature overlap is almost similar for both BERT and XLM-RoBERTa, we chose the one with the better performance on test data (BERT in our case). The ensemble model improved the overall performance over baselines (approx. 30%), though the relative improvement over individual constituents is not significant. We probed further into the performance of ensemble method. We observed that ensembles are useful specially in the cases where sentence is ambiguous and doesn't have a straight-forward sentiment. Considering a tweet from the test dataset which says "i cant stay away from bug thats my baby", the BERT model predicted this tweet as positive with 0.94 probability whereas RoBerta predicted it as negative with 0.96 probability[1]. One can clearly see the ambiguous nature of this tweet as both the models are extremely confident in their respective predictions. Taking ensemble reduces this bias by taking both models into account and thus the averaging would classify this tweet as negative by 0.51 probability. Note that such disagreement between BERT and RoBERTa is observed only for a restrictive subset of ambiguous tweets and we will see next in Figure 3 that both models agree in general.

IV. Analysis and Discussion

In this section, we present an empirical analysis of the functioning of our proposed method, since it is equally important to interpret the predictions as much as appreciate their accuracy. Since the three methods we use work for individual models, we choose to scrutinize the better performing component of the ensemble, namely RoBERTa.

[1]logit values can be found in exps/bert-full-data/bert-full-data.logits.csv and exps/roberta-cust/roberta-cust.logits.csv (Id=3)



(a) BERT (b) RoBERTa (c) XLM-RoBERTa (d) XLNet

Figure 2: Visualization of the learnt features from different models in 2 dimensions using Isomap. The blue points in the plot denote the positive sentiments while the red one denote the negative ones.

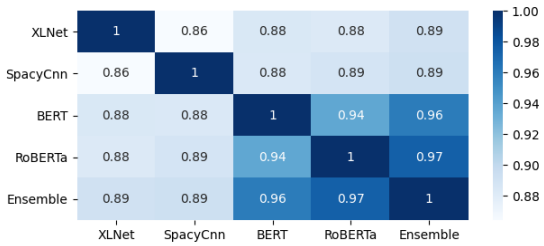


Figure 3: Pairwise Agreement on test set prediction between different models. Transformer-based models tend to agree more among each other.

Attention mechanism lies at the core of our models, hence it becomes indispensable to analyse how our model attends to parts of a tweet. We visualize the attention scores for each token of a given tweet in all the layers of the model using the tool BertVis [31]. Figure 4(a) shows the attention contribution of various sentence tokens (on right) for predicting the token "pain" (on left). Higher intensity lines indicate larger attention values. Figure 4(a) depicts the first layer of the model in which all the tokens have significant contribution, whereas Figure 4(b) shows attention contribution in the last layer (layer 11), where only tokens in close proximity attend to *pain*. This analysis shows that initial layers keep a broader

attention span which gets fine tuned to only the important nearby tokens in the subsequent layers.

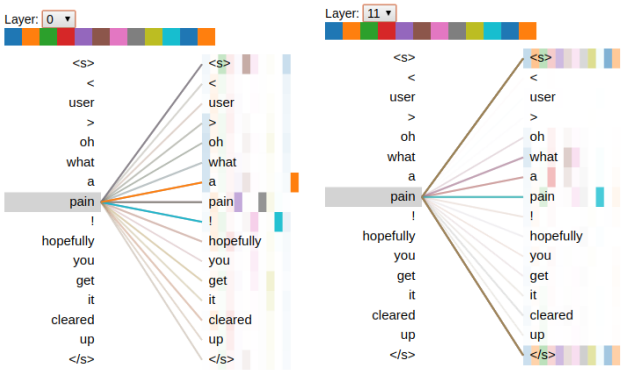
We next proceed to analyse which subtext of a tweet is most likely responsible for the prediction of the sentiment. We use a different tweet dataset from the Kaggle challenge: Text Sentiment Extraction[2]. In order to predict subtext indices, we trained additional regression units on top of features extracted from RoBERTa. We then selected some test samples and assumed both positive and negative sentiments for them. Table II shows the subtext with the largest contribution to the overall sentiment under both assumptions. Not just that the subtexts identified for positive and negative scenarios are in sharp contrast to each other in literal meaning, they also very well align with the general connotation expressed by their respective sentiment classes. This shows that RoBERTa adapts its attention scores over the tokens based on the given sentiment in an efficient and robust manner.

We further investigate the impact of not just the best one, but all significant words on the overall sentiment of a tweet. We also scrutinize the degree of contribution of each such word, even when they convey conflicting sentiments. We use Local Interpretable Model-agnostic Explanations, LIME [32] in order to interpret the prediction of our proposed model. LIME repeatedly masks different words of a text and estimates their contribution towards predicting the sentiment as shown in figure 5. In the example shown, the tweet above has a ground truth value of positive sentiment, similar to the model's prediction. The intensity of the color represents how strongly the word contributes to the overall sentiment of the tweet. The model majorly focuses on *good* in order to make prediction, however it does consider *lostt*, conveying a negative sentiment, as a promising candidate too. Similarly, the model primarily focuses on words with negative connotation in the second example, and classifies it as a negative tweet. This exercise shows that our proposed model is quite robust to the presence of words with conflicting intentions withing the same tweet and learns to selectively attend to the ones most relevant to the overall sentiment conveyed.

Tweet text	Positive	Negative
<user>oh what a pain ! hopefully you get it cleared up	hopefully	pain
<user>thankyou sooo much for letting me stress out haha xxx	sooo thankyou much	out stress
<user><user>nice unfortunately i don't get podcasts have you checked out techi ?	nice	unfortunately

Table II: Subtexts detected with positive and negative sentiments on the same text.

[2]https://www.kaggle.com/c/tweet-sentiment-extraction



(a) layer 0 (b) layer 11

Figure 4: Visualization of attention scores of various token for predicting the token "pain" in layer 0 and layer 11 of RoBERTa model (color palate on the top indicates all attention heads are combined).

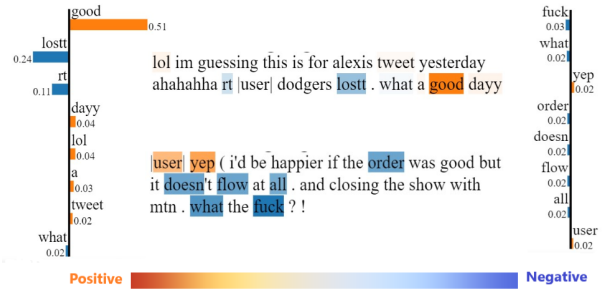


Figure 5: LIME interpretation of model prediction on tweets with positive (above) and negative (below) sentiments. Our model robustly focuses on words which convey the sentiment similar to the ground-truth.

V. Conclusion

In this work we present a series of experiments which help us understand what the transformer architectures are learning in an interpretable way. We propose an ensemble of BERT and RoBERTa architectures to predict the tweet sentiments. We critically assessed each stage, investigating and reinforcing the observations. Overall our proposed model performs fairly well and we finish in top 5 on the public leaderboard.

At the end, we thank Data Analytics Lab at ETH Zurich for their constant support during the semester, SIS HPC team at ETH Zurich for the well-maintained Leonhard server and finally the Hugging face [33] team for creating such a streamlined library which made the experimentation effortless and uncomplicated.

REFERENCES

[1] B. Liu and L. Zhang, "A survey of opinion mining and sentiment analysis," in *Mining text data*. Springer, 2012, pp. 415–463.

[2] A. Giachanou and F. Crestani, "Like it or not: A survey of twitter sentiment analysis methods," *ACM Computing Surveys (CSUR)*, vol. 49, no. 2, pp. 1–41, 2016.

[3] H. Saif, Y. He, and H. Alani, "Alleviating data sparsity for twitter sentiment analysis." CEUR Workshop Proceedings (CEUR-WS. org), 2012.

[4] E. Kouloumpis, T. Wilson, and J. Moore, "Twitter sentiment analysis: The good the bad and the omg!" in *Fifth International AAAI conference on weblogs and social media*, 2011.

[5] S. Kiritchenko, X. Zhu, and S. M. Mohammad, "Sentiment analysis of short informal texts," *Journal of Artificial Intelligence Research*, vol. 50, pp. 723–762, 2014.

[6] H. Saif, M. Fernández, Y. He, and H. Alani, "On stopwords, filtering and data sparsity for sentiment analysis of twitter," 2014.

[7] O. Owoputi, B. O'Connor, C. Dyer, K. Gimpel, N. Schneider, and N. A. Smith, "Improved part-of-speech tagging for online conversational text with word clusters," in *Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: human language technologies*, 2013, pp. 380–390.

[8] S. Narr, M. Hulfenhaus, and S. Albayrak, "Language-independent twitter sentiment analysis," *Knowledge discovery and machine learning (KDML), LWA*, pp. 12–14, 2012.

[9] A. Severyn and A. Moschitti, "Twitter sentiment analysis with deep convolutional neural networks," in *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2015, pp. 959–962.

[10] Y. Wang, M. Huang, X. Zhu, and L. Zhao, "Attention-based lstm for aspect-level sentiment classification," in *Proceedings of the 2016 conference on empirical methods in natural language processing*, 2016, pp. 606–615.

[11] R. Socher, C. C. Lin, C. Manning, and A. Y. Ng, "Parsing natural scenes and natural language with recursive neural networks," in *Proceedings of the 28th international conference on machine learning (ICML-11)*, 2011, pp. 129–136.

[12] R. Socher, J. Pennington, E. H. Huang, A. Y. Ng, and C. D. Manning, "Semi-supervised recursive autoencoders for predicting sentiment distributions," in *Proceedings of the 2011 conference on empirical methods in natural language processing*, 2011, pp. 151–161.

[13] L. Dong, F. Wei, C. Tan, D. Tang, M. Zhou, and K. Xu, "Adaptive recursive neural network for target-dependent twitter sentiment classification," in *Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 2: Short papers)*, 2014, pp. 49–54.

[14] D. Tang, B. Qin, X. Feng, and T. Liu, "Effective lstms for target-dependent sentiment classification," *arXiv preprint arXiv:1512.01100*, 2015.

[15] M. Zhang, Y. Zhang, and D.-T. Vo, "Gated neural networks for targeted sentiment analysis," in *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[16] S. Ruder, P. Ghaffari, and J. G. Breslin, "A hierarchical model of reviews for aspect-based sentiment analysis," *arXiv preprint arXiv:1609.02745*, 2016.

[17] D. Ma, S. Li, X. Zhang, and H. Wang, "Interactive attention networks for aspect-level sentiment classification," *arXiv preprint arXiv:1709.00893*, 2017.

[18] B. Huang and K. M. Carley, "Parameterized convolutional neural networks for aspect level sentiment classification," *arXiv preprint arXiv:1909.06276*, 2019.

[19] X. Li, L. Bing, W. Lam, and B. Shi, "Transformation networks for target-oriented sentiment classification," *arXiv preprint arXiv:1805.01086*, 2018.

[20] D. Tang, B. Qin, and T. Liu, "Aspect level sentiment classification with deep memory network," *arXiv preprint arXiv:1605.08900*, 2016.

[21] P. Chen, Z. Sun, L. Bing, and W. Yang, "Recurrent attention network on memory for aspect sentiment analysis," in *Proceedings of the 2017 conference on empirical methods in natural language processing*, 2017, pp. 452–461.

[22] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to fine-tune BERT for text classification?" *CoRR*, vol. abs/1905.05583, 2019. [Online]. Available: http://arxiv.org/abs/1905.05583

[23] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized BERT pretraining approach," *CoRR*, vol. abs/1907.11692, 2019. [Online]. Available: http://arxiv.org/abs/1907.11692

[24] Z. Yang, Z. Dai, Y. Yang, J. G. Carbonell, R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," *CoRR*, vol. abs/1906.08237, 2019. [Online]. Available: http://arxiv.org/abs/1906.08237

[25] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," 2019.

[26] Z. Dai, Z. Yang, Y. Yang, J. G. Carbonell, Q. V. Le, and R. Salakhutdinov, "Transformer-xl: Attentive language models beyond a fixed-length context," *CoRR*, vol. abs/1901.02860, 2019. [Online]. Available: http://arxiv.org/abs/1901.02860

[27] "Python library for word segmentation: http://www.grantjenks.com/docs/wordsegment/."

[28] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543. [Online]. Available: http://www.aclweb.org/anthology/D14-1162

[29] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California: Association for Computational Linguistics, Jun. 2016, pp. 1480–1489. [Online]. Available: https://www.aclweb.org/anthology/N16-1174

[30] M. Honnibal and I. Montani, "spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing," 2017, to appear.

[31] J. Vig, "A multiscale visualization of attention in the transformer model," 01 2019, pp. 37–42.

[32] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should I trust you?": Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, 2016, pp. 1135–1144.

[33] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, and J. Brew, "Huggingface's transformers: State-of-the-art natural language processing," *ArXiv*, vol. abs/1910.03771, 2019.

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Declaration of originality

The signed declaration of originality is a component of every semester paper, Bachelor's thesis, Master's thesis and any other degree paper undertaken during the course of studies, including the respective electronic versions.

Lecturers may also require a declaration of originality for other written papers compiled for their courses.

I hereby confirm that I am the sole author of the written work here enclosed and that I have compiled it in my own words. Parts excepted are corrections of form and content by the supervisor.

Title of work (in block letters):

CIL: TEXT SENTIMENT CLASSIFICATION

Authored by (in block letters):

For papers written by groups the names of all authors are required.

Name(s):	**First name(s):**
SHEIKH	NOMAN
SINGH	RISHABH
JAIN	SAMRIDDHI
MOHIT	KUMAR

With my signature I confirm that

- I have committed none of the forms of plagiarism described in the 'Citation etiquette' information sheet.
- I have documented all methods, data and processes truthfully.
- I have not manipulated any data.
- I have mentioned all persons who were significant facilitators of the work.

I am aware that the work may be screened electronically for plagiarism.

Place, date

ZURICH, 31st JULY 2020

Signature(s)

DocuSigned by:
Noman Ahmed Sheikh
DocuSigned by:
737E404EADD24E3...
Rishabh Singh
B10138ACE9164F0...
DocuSigned by:
samriddhi jain
DocuSigned by:
23847027808C4BF...
Kumar Mohit
6F061285475D449...

For papers written by groups the names of all authors are required. Their signatures collectively guarantee the entire content of the written paper.