ITM Web of Conferences **56**, 04007 (2023) https://doi.org/10.1051/itmconf/20235604007
ICDSAC 2023

Hate Speech Detection in Twitter Using Different Models

Anagha Abraham
Department Of Computer Science and Technology
Rajagiri School Of Engineering And Technology
Kochi, India
anaghaabraham2000@gmail.com

Antony J Kolanchery
Department Of Computer Science and Technology
Rajagiri School Of Engineering And Technology
Kochi,India
antonyjkolenchery@gmail.com

Anugraha Antoo Kanjookaran
Department Of Computer Science and Technology
Rajagiri School Of Engineering And Technology
Kochi, India
anugrahaantoo@gmail.com

Binil Tom Jose
Department Of Computer Science and Technology
Rajagiri School Of Engineering And Technology
Kochi, India
biniltomjose12780@gmail.com

Dr Dhanya PM
Department Of Computer Science and Technology
Rajagiri School Of Engineering And Technology
Kochi, India
dhanya_pm@rajagiritech.edu.in

***Abstract*—Twitter's primary objective is to facilitate free expression and the exchange of ideas, allowing individuals to share their thoughts, opinions, and information with others without any limitations or constraints. It helps a human being to perceive different scopes and points of view. It is used to serve the public discussion and it should not be used to undermine individuals based on their race, nationality, public standing, rank, sexual orientation, age, disability, or health conditions. So, using hate speech is not appropriate and removal of hate speech is necessary for achieving the goal. This paper aims to utilize machine learning algorithms such as Logistic Regression, Support Vector Machine, Random Forest, CNN-LSTM, and Fuzzy method to compare and evaluate their accuracy in detecting hate speech. The objective is to determine the best model for hate speech detection.**

***Index Terms*—SVM, Random Forest, Logistic Regression, CNN-LSTM, Fuzzy Classification**

I. Introduction

Hate speech is any content that generates or manipulates an individual or community based on qualities such as colour, ethnicity, gender, sexuality, nationality, or religion. The usage of social media in daily life has expanded in such a way that everyone has the space to think and write anything they want. The prevalence of hate speech has been increasing day by day, making it necessary to automate detection of hate speech. We have used machine learning approaches to streamline the classification process of identifying hate speech within Twitter data. These approaches simplifies the classification process, making it more efficient and effective in identifying and detecting hate speech. The rise in social media usage has led to an increase in hate speech being disseminated on these platforms. With the sheer volume of content being shared, it has become impractical to manually identify instances of hate speech. Therefore, the development of an automated hate speech detection model is crucial. This research delves into various Natural Language Processing approaches for the purpose of hate speech classification using Machine Learning algorithms. The objective of this research is to provide an effective tool for the automated detection and categorization of hate speech in social media platforms. The study uses a publicly available dataset on hate speech from Twitter. The research work compares multiple classification approaches that are based on distinct document representation strategies and text classification models. Hence this research paper mainly focus on text classification purpose that will help to identify the hate speech in twitter using natural language processing. Using the best training model the hate speech will be detected.

II. Related Works

A. The various model used to classify hate speech detection are:

- Support Vector Machine(SVM) :Support vector machine, also known as SVM, is a popular machine learning technique for both classification and regression tasks. It is a supervised learning algorithm that can identify patterns in data and classify new data based on those patterns. However, primarily, it is used for Classification problems in Machine Learning. It mainly focus on choosing the vectors in creating the hyperplanes, and thereby called as support vectors.

It's main focus is to find and search the decision boundary that helps in separating the data points into different classes. SVM represents each data point as a vector in a high-dimensional space, and it attempts to identify a hyperplane that can effectively separate the data points into different classes. The method finds the hyperplane with the shortest distance between itself and the closest data points in each class.

When it is linearly inseparable data and when small number of training examples are used, the SVM comes into play and it is also much less prone to overfitting as compared to other algorithms as it tries to maximize the margin between the decision boundary and the data points.

In conclusion, SVM can be used for both classification and regression that makes it a powerful and flexible machine learning algorithm.Hence, SVM is used in this project to classify the hate speech and the non-hate speeches such in a way that can handle high-dimensional data with a relatively small number of training examples.

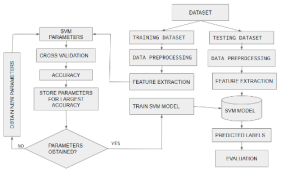


Fig. 1. Archiectural Diagram SVM

- Logistic Regression: Logistic Regression is a prominent machine learning algorithm that is often employed for a variety of classification tasks, including hate speech detection in social media platforms like Twitter. In this approach, each tweet is represented as a set of features, such as the frequency of certain keywords, the length of the tweet, and the sentiment of the tweet and these features are then used to train a logistic regression classifier, which can predict then be used to check whether a tweet is hate speech or not.

  Logistic regression can be implemented for hate speech detection in Twitter using the following steps:

  – Data preprocessing: The stop words are eliminated from the Twitter dataset followed by stemming and lowercasing every word.
  – Feature extraction: Data can be vectorized using term frequency-inverse document frequency (TF-IDF).
  – Labeling the data: Label whether the data is hate speech or not, or a labeled dataset can be used.
  – Train the model: Train a logistic regression classifier on the labeled data using the extracted features.
  – Evaluate the model: Analyse the model's performance using criteria like accuracy, recall, F1 score and precision on a test dataset.
  – Fine-tune the model: The model is to fine-tuned by adjusting the hyperparameters of the logistic regression algorithm and feature extraction techniques.

  Overall, logistic regression is a simple and effective approach for hate speech detection in Twitter, especially when working with limited data. However, it may not be able to capture complex relationships between features, and more sophisticated models such as neural networks may be needed for improved performance.

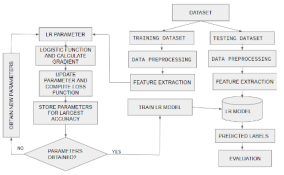


Fig. 2. Archiectural Diagram LR

- Random Forest: Random Forest is an ensemble learning method that is used for classification task in machine learning. In this method multiple decision trees are combined to improve the accuracy and robustness of the model. In Random Forest, decision trees are trained using randomly chosen training data subsets. Each decision tree in the forest makes a prediction for a given input, and the final prediction is determined by taking the average of the predictions of all the trees. This helps to reduce the variance and overfitting issues that can arise with individual decision trees.

  Random Forest works by randomly selecting subsets of the training data and features for each decision tree. This is done so as to ensure that each tree in the forest is trained on a slightly different subset of the data, which helps to reduce the correlation between the trees. The algorithm also uses a technique called bagging (bootstrap aggregating) to randomly sample subsets of the data with replacement, which helps to improve the stability of the model.

  Random Forest outperforms other machine learning methods in various ways. It can work with both categorical and continuous input data and can perform regression and classification tasks, making it a popular choice for a wide range of applications. It is also robust to overfitting and can handle missing values in the input data. Moreover, Random Forest is computationally efficient and can handle high-dimensional data with a large number of features.

  In conclusion, Random Forest is a powerful ensemble learning method which can be used for hate speech detection in Twitter that combines multiple decision trees to improve the accuracy and robustness of the model. To lessen the correlation between the trees and increase the stability of the model, it chooses subsets of the training data and features at random for each decision tree.

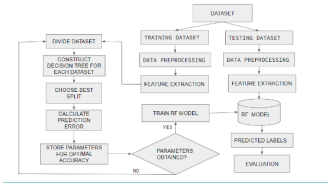


Fig. 3. Archiectural Diagram Random Forest

- CNN-LSTM: It is a deep learning architecture method that combines Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks to detect patterns and connection present in sequential data. The CNN component of the architecture is used to extract local features from the input data. For text-based input data like tweets, the CNN may use filters of different sizes to capture n-grams (i.e., sequences of n words) that are typical of hate speech. The output of the CNN consist of a set of feature maps which helps to represent the presence of a particular n-grams present in the output. The LSTM component of the architecture captures the global context and relationships between the local features extracted by the CNN. LSTM is a recurrent neural network that remembers information present from the previous steps and it also helps to remove the information that is no longer relevant. This makes LSTM to capture long-term dependencies in the input data which helps to detect whether hate speech is present or not.
  The output of LSTM actually represents hiddens states which helps to represent the context present in a particular time step. The hidden states are then concatenated and passed through a fully connected layer to make the final prediction of whether the input contains hate speech or not.
  The CNN-LSTM architecture has several supremacy over other deep learning architectures. The CNN captures local features that are important for hate speech detection, while the LSTM can capture long-term dependencies that are crucial for understanding the context of the input data. By combining both CNN and LSTM we would be able to achieve high accuracy in hate speech detection tasks.

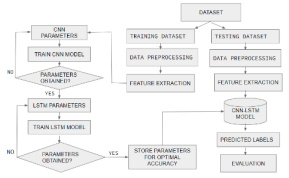


Fig. 4. Archiectural Diagram CNN-LSTM

- Fuzzy Classification: In Twitter hate speech detection, fuzzy classification has been applied as a powerful machine learning technique. Unlike traditional machine learning algorithms, it handles the uncertainty and subjectivity of human language. By using Natural Language Processing (NLP), classifiers can automatically analyze text and then assign a set of pre-defined tags or categories based on its content and hence can be used in assigning tags to the hyperclass. In fuzzy classification, membership grades are assigned to each class, which represent the degree of a data point's membership. Class membership grades are computed using similar features of data points and class features. Class membership grades are typically represented as values between 0 and 1, where 0 indicates the data point does not belong to the class at all, and 1 indicates it definitely does.
  The first step in implementing fuzzy classification for hate speech detection on Twitter is to remove stop words, stem, and lowercase the text.The next steps involves extracting features from the preprocessed data using methods like term frequency-inverse document frequency (TF-IDF) and n grams.
  Data needs to be labeled as containing hate speech after feature extraction,this can be done using a labeled dataset that has been manually annotated by human annotators.
  Using the extracted features, we train a fuzzy classifier on the labeled data. According to the input data, the fuzzy classifier assigns membership grades to each class using fuzzy logic rules,the classifier can be fine-tuned by adjusting the parameters that control the degree of fuzziness and the threshold for assigning membership grades.
  A test dataset can also be used to evaluate fuzzy classifier performance based on metrics such as recall, accuracy, and precision,these metrics provide a measure of how well the classifier can identify hate speech in Twitter data. Fuzzy classification can handle human language's inherent ambiguity and subjectivity, making it an ideal technique for hate speech detection on Twitter. By assigning membership grades to each class, it can provide more nuanced predictions and better handle borderline cases of hate speech.

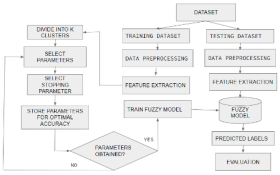


Fig. 5. Architectural Diagram Fuzzy

III. FUZZY CLASSIFICATION

A fuzzy classification allows objects to belong to multiple categories or classes simultaneously with varying degrees of membership.Contrary to the classic crisp classification system,

in which objects are assigned to a single class based on a strict set of rules or criteria, we are using a fuzzy classification system.

An object is evaluated in fuzzy classification according to a set of criteria or features, and is then assigned a degree of membership based on a function that determines its relative belongingness to each class according to a set of criteria or features.It is important to note that the degree of membership can also range from 0 to 1, where 0 is the level of no membership and 1 represents the level of full membership.

Fuzzy classifications can be useful not only when dealing with complex and ambiguous data, but they have also proven useful when dealing with objects' characteristics that are difficult to categorize in a sharp, binary way, such as those relating to their location or size.There are many fields where it can be applied, such as pattern recognition, machine learning, and artificial intelligence.

In a few different ways, fuzzy classifications can be helpful in detecting hate speech in a variety of different circumstances. These are as follows:

- Language ambiguity: Hate speech is commonly ambiguous, or has multiple meanings, so it can be difficult to classify using traditional binary classification methods for determining its classification.Using fuzzy classification can help account for this ambiguity by offering more nuanced classifications.
- Data that is imbalanced: It appears that hate speech detection often involves datasets that are imbalanced, in the sense that there is a much smaller number of instances of hate speech than not-hate speech.As a result of fuzzy classification, decision boundaries can be more flexible, allowing for better data capture.
- Multi-feature analysis: Hate speech detection often takes into account sentiment and language use in the text.It has been suggested that fuzzy classification is a useful tool to integrate these features together, allowing a more holistic analysis of hate speech to be performed.
- Uncertainty: Hate speech detection inherently involves uncertainty, as disagreements or ambiguities may arise regarding what constitutes hate speech, which could lead to incorrect results.Using fuzzy classification can help account for this uncertainty by allowing classifications with varying degrees of confidence.

In general, fuzzy classification can be an extremely powerful tool in the detection of hate speech and other forms of harmful language since it is able to provide more nuanced and flexible classifications that are able to capture the complexity of the hate speech and other types of harmful language.

IV. FUZZY HYPERGRAPH

Fuzzy hypergraph can be extended from the usual traditional hypergraph that helps in pointing out the relationship between the objects and their relations. In fuzzy hypergraph, membership degree is assigned to each object and which helps in identifying the involovent of the degree of object in the hyperedge. Fuzzy hypergraph has been applied to the task of hate speech detection in Twitter to handle the ambiguity and uncertainty that is inherent in human language. The goal is to represent the relationships between the Twitter data in a way that captures the nuances of language and the complex interactions between different elements of the data.

To implement fuzzy hypergraph for hate speech detection in Twitter, firstly, data preprocessing steps are done later the data is represented as nodes and then relationship between the nodes are shown through the fuzzy hypergraphs.

The next step is to construct a hypergraph from the nodes and their relationships. In fuzzy hypergraph, each hyperedge is assigned a membership grade to each node, representing the degree of involvement of the node in the hyperedge. The membership grades are typically represented as values between 0 and 1, where a value of 0 means the node does not belong to the hyperedge at all, and a value of 1 means the node definitely belongs to the hyperedge.

The construction of hypergraph can be done using using algorithms, such as the clustering algorithm or the association rule mining algorithm. The specific problem and the characteristics of the data helps in choosing which algorithm to be used.

After the hypergraph is constructed, detection of hate speech can be done by training a fuzzy classifier. The classifier may use fuzzy logic rules to assign membership grades to each class by using n the features of the hypergraph. In Natural Language Processing (NLP) Hypergraph is used to represent complex relationships between words, phrases, and concepts.

By allowing edges to connect more than two vertices,a hypergraph is a mathematical structure that generalizes the notion of a graphs. In NLP, hypergraphs can be used to represent various linguistic structures, such as syntax, semantics, and discourse.

One application of hypergraphs in NLP is in semantic parsing, where the goal is to map natural language sentences to their corresponding formal meaning representations. Along with the semantic information associated with each word or phrase a hypergraph can be used to represent the syntactic structure of a sentence, whereas the main goal is to model the relationships between different concepts in a knowledge base. Hypergraphs can be used to represent complex relationships between entities, such as in the case of knowledge graphs, which are used to represent large-scale semantic networks.

Overall, hypergraphs provide a flexible and powerful tool for modeling complex relationships in NLP, and have many potential applications in areas such as machine translation, sentiment analysis, and information extraction.

Finally, accuracy, recall, F1 score and precision are use dto analyse the performance of the classifier on a test dataset. These metrics shows how accurately the fuzzy classifier can identify and detect the dataset in Twitter.

In total, fuzzy hypergraph can be used to capture the complex relationships between words and phrases in the data. It can provide more accurate and correct prediction with the help of assigning membership grades to each of the hyperedges and can handle borderline cases of hate speech.

A. Weight Assigning

Model words as nodes and edges as sentences and assign weights.

Let H= (Hn, He) be a neutrosophic hypergraph where Hn and He be the family of nodes and hyperedges. For each node n in H.; TA(n) € [0, 1], IA(n) E [0, 1],FA(n) € [0, 1] and TA(n) + IA(n) + FA(n)= 3, where TA(n) is the truth membership function, IA(n) is the indeterminacy membership function and FA(n) is the falsity membership function.

- CASES INCLUDED:
- Case 1: [0,0,1] - Word is not a hate word.
- Case 2: [1,0,0] - Sure Hate word.
- Case 3: [1, 0.5,0] - Hate word, but depends on context.
- Case 4: [0.5, 1, 0] - possibility of being a hate word, but high indeterminacy.
- Case 5: [0, 0.5, 1] - Not hate word, but depends on context.

Similarly for each edge e in He, TA(e) E [0, 1], IA(e) E [0, 1], FA(e) E [0, 1]. and TA(e) + IA(e) + FA(e) ≤ 3,

where TA(e) is the truth membership function, IA(e) is the indeterminacy membership function and FA(e) is the falsity membership function.

B. MEMBERSHIP DEGREE

The edge membership degree (TA(e), IA(e), FA(e)) is defined as follows and is given by :

- TA (e) = Max(TA(n)); Vn € e
- IA (e) = avg(IA(n)); Vn € e
- FA (e) = avg(FA(n)); Vn € e

C. SUB HYPERGRAPH

- Apply (a, B, y) cut on the hypergraph created, such that a sub hypergraph X is created with (TA(e), IA(e), FA(e)) = [0.5,0.3,0]
- Apply higher level (a, SS, y) cut on the hypergraph created, such that a sub hypergraph A is created with (TA(e), IA(e), FA(e)) = [0.8,0.3,0]

D. Morphologocal Operations On Subgraph

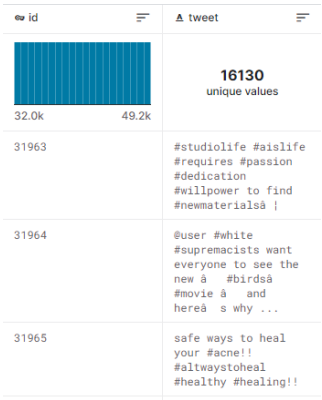
- Dilation w.r.t Nodes and Edges: Dilation is a common operation in fuzzy hypergraph theory that involves expanding the neighborhood of a node in a hypergraph.As a result, when a node in a fuzzy hypergraph is dilated, new nodes are added based on how similar they are to the initial node.
  Identifying nodes that are similar to the original node in a fuzzy hypergraph can be done using similarity measures such as the Jaccard index or cosine similarity, the degree of similarity can then be used to determine how strongly the new nodes should be connected to the original node and to other nodes in the hypergraph. δa(X)-dilation of X w.r.t nodes, is the set of all words in X. Edge dilation in a fuzzy hypergraph involves adding new edges to the hypergraph based on the degree of similarity between edges ,this operation is also known as edge expansion or edge growth. To perform edge dilation in a fuzzy hypergraph, one can use a similarity measure such as the Jaccard index or cosine similarity to identify edges that are similar to each other. Based on similarity, the hypergraph can be constructed by strongly connecting the new edges to the existing edges. δ(X)-dilation of X, w.r.t edges, is the set of all sentences consisting of nodes in X.
- Erosion w.r.t Nodes and Edges: The erosion of a fuzzy hypergraph involves removing nodes that are not similar enough to the original node in order to shrink the neighborhood, this operation is also known as node contraction or node pruning.
  When performing erosion with respect to a node within a fuzzy hypergraph, one can identify nodes that are not similar enough to the original node using a similarity measure like the Jaccard index. The degree of similarity can then be used to determine which nodes should be removed from the hypergraph. En(X) is erosion of X w.r.to nodes, which is the set of all nodes in X and not in X'. In a fuzzy hypergraph, erosion involves removing edges that are not similar enough to each other to shrink the edges, this operation is also known as edge contraction or edge pruning.
  One can perform erosion within a fuzzy hypergraph by identifying edges that are not similar enough to each other using a similarity measure such as the Jaccard index, the degree of similarity can then be used to determine which edges should be removed from the hypergraph. E(X) - is erosion of X w.r.to edges, which is the set of all edges consisting of X" only.

E. Hate Speech Removal

- Skelton Operation: Apply skelton operation S(HOne can perform erosion within a fuzzy hypergraph by identifying edges that are not similar enough to each other using a similarity measure such as the Jaccard index, = H- (Se (Xn)) k. This removes the hate speech from the tweets and retain the non-hate sentences.
- Miss or Hit Operation: Apply Hit or Miss operation as Hit-or-Miss (H,A) = A EXnA' € (W-X) where W is the window of X obtained by dilating X. This intersection will give some acute hate words. i. Apply thinning operation as ii. H-Se (Hit-or-Miss(H,A))
  The resultant sentences of the above two operations give tweets avoiding hate speech.

F. IMPLEMENTATION

- DATASET: The dataset for the hatespeech detection using various classifying technique has been mainly divided into four categories. The four different categories are Bodyshamming, Ethical, Suicidal, Aggressive. The dataset below shown provides an example for these categories:
- Comparative analysis: The hate speech detection in twitter has been analysed using five classifier. Comparative Analysis shows that SVM is more accurate in detecting the hate speech than the other classifiers like Random Forest, logistic regression, CNN-LSTM,



id	tweet
32.0k 49.2k	16130 unique values
31963	#studiolife #aislife #requires #passion #dedication #willpower to find #newmaterialsâ ¦
31964	@user #white #supremacists want everyone to see the new â #birdsâ #movie â and hereâ s why ...
31965	safe ways to heal your #acne!! #altwaystoheal #healthy #healing!!

Fig. 6. WLCG

FUZZY classifiers. Support Vector Machine shows an accuracy of 88.69, Random forest shows an accuracy of 87.88, logistic regression shows an accuracy of 86.55,Fuzzy shows 75.55 and the fuzzy hypergraph shows the accuracy of 75.35.The fuzzy hypergraph is so essential for the classification using fuzzy classifier, it also helps or focus in diagramtically showing how each words in a sentence are connected to each other and its important in determining the membership degree function and its relation. fuzzy hypergraph helps in doing the morphological operations such as dilation and erosion and helps in determing the connection and to perform the process of removing the hate speech detection from social platforms such as twitter. Support vector machine has been selected to predict hate speech on Twitter in accordance with the trained model's accuracy prediction.

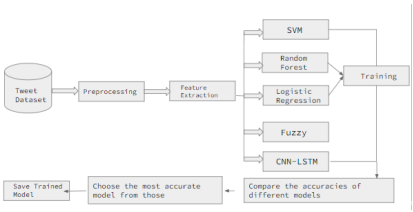


Fig. 7. WLCG

G. CONCLUSION

In Conclusion, the project on hate speech detection in Twitter mainly focus to shed light on the importance of developing effective solutions to tackle the problem of hate speech in social media. Through the application of machine learning and natural language processing techniques, the project has explored various approaches, including Fuzzy classification, Random Forest, logistic regression, CNN-LSTM and SVM.

The project findings suggest that detecting hate speech in Twitter these machine learning approaches can be used effectively. However, the accuracy of the models depends on The quality of the training data and the algorithm's ability to understand helps in determining the accuracy of the model.The main focus was to classify using different models and then predict the hatespeeches using the best model with more accuracy values . Fuzzy classifier is one of the significant method that helps in classifying these hatespeeches by creating the fuzzy hypergraphs and then by providing the membership degree values.

The project highlights the significance of addressing hate speech in social media platforms to ensure they remain inclusive and safe for all users. Furthermore, it showcases the potential of natural language processing and machine learning techniques in developing effective solutions to tackle hate speech in social media.

Moving forward, Inorder to enhance the accuracy and robustness of these models furthur study will be conduced to detect various types of hate speech. In conclusion, this project has demonstrated the importance of addressing the issue of hate speech in social media and the potential of technology in developing effective solutions to tackle this problem

REFERENCES

[1] Razmi, N.A., Zamri, M.Z., Ghazalli, S.S.S. and Seman, N., 2021. Visualizing stemming techniques on online news articles text analytics. Bulletin of Electrical Engineering and Informatics, 10(1), pp.365-373.

[2] Hickman, L., Thapa, S., Tay, L., Cao, M. and Srinivasan, P., 2022. Text preprocessing for text mining in organizational research: Review and recommendations. Organizational Research Methods, 25(1), pp.114-146.

[3] Liu, Q., Wang, J., Zhang, D., Yang, Y. and Wang, N., 2018, December. Text features extraction based on TF-IDF associating semantic. In 2018 IEEE 4th International Conference on Computer and Communications (ICCC) (pp. 2338-2343). IEEE.

[4] Akuma, S., Lubem, T. and Adom, I.T., 2022. Comparing Bag of Words and TF-IDF with different models for hate speech detection from live tweets. International Journal of Information Technology, pp.1-7.

[5] Han, K.X., Chien, W., Chiu, C.C. and Cheng, Y.T., 2020. Application of support vector machine (SVM) in the sentiment analysis of twitter dataset. Applied Sciences, 10(3), p.1125.

[6] Guo, X., Tian, B. and Tian, X., 2022. HFGNN-Proto: Hesitant Fuzzy Graph Neural Network-Based Prototypical Network for Few-Shot Text Classification. Electronics, 11(15), p.2423.

[7] Lucie Flek Mainz University of Applied Sciences Germany. Returning the N to NLP: Towards Contextually Personalized Classification Models

[8] Hui Liu1, Qingyu Yin2, William Yang Wang3. Towards Explainable NLP: A Generative Explanation Framework for Text Classification