ANALYSIS OF ONLINE HATE SPEECH AND THE APPLICATION OF CLASSIFICATION MODELS IN DETECTING ONLINE HATE SPEECH

Submitted by:
Ami, Julius Lorenz
Capulong, Mark
Cueva, Larry Miguel
Osorio, Bryan
Ore, Gabriel
Panlilio, Andrei Shane
Zabala, Leonard Howell

Submitted to:
Prof. Montaigne Molejon
CS Elective 4 - Data Mining

Table of Contents

List of Figures

Description of the Problem

Freedom of speech or expression is very important for a country to give everyone a right to express their thoughts and opinions about a certain idea or topic. It is defined as the right to speak, write and express an individual or group opinion without facing punishment. But not every thought or opinion they express is a sensible one, some might be below the belt and directed towards another person or group of people, and even outside the topic that is being discussed. This is where hate speech comes along, it is hate speech when it is directly attacking a person or group of people in an offensive manner.

In order to encourage strategies for countering online hate speech, we introduce a novel task of generative hate speech intervention along with two fully-labeled datasets collected from Gab and Reddit. Due to the rise of social media use nowadays, there are a lot of people exchanging various opinions that can be positive or negative, thus the rise in hate speech.

Hate speech is one of, if not the most, common, sometimes funny, but most times damaging objects on the Internet manufactured by different kinds of people using social media sites. Brown (2017) defined hate speech as any textual or verbal practice that implicates the issue of discrimination or violence against people in regard to the race, religion, ethnicity, nationality, sexual orientation, and gender identity of an individual. It results in an unhealthy environment of exchange online, and encourages activity and morality that is, in many terms, uncivilized, hateful, and discriminatory – just to name a few. There are a lot of countermeasures taken to combat this daily frenzy of hate speech online, but even the most advanced algorithms are either still being evaded by hate speech doers, or are too aggressive as algorithms that even the slightest sign of hate or vulgarity, that is with a reasonable and justifiable context to the cause of inception, immediately leads to bans, penalties or censorship.

Increasing problems in the usage and existence of hate speech in this day and age of social media have become a pain for the platforms themselves to handle. People from all around the globe use these platforms on a daily basis, and the number of new users only increases steadily day by day. With that thought in mind, it can be said that all the hate speech one sees on their timeline is just the tip of the iceberg, to a whole sea of the Internet's entirety of hate speech. Mitigating the problem in its infestation of the online world has always been the objective of most social media algorithms, and they are, to this day, still in the process of constantly improving their detection algorithms to better address the issue at hand. This gave the proponents an idea to dive in on finding solutions to the problem and create a hate speech classifier that will be able to classify hate speeches and create a deep analysis of hate speech.

Literature Review

The study by Qian, et al. in 2019 discusses countering online hate speech is a critical yet challenging task, but one which can be aided by the use of Natural Language Processing (NLP) techniques. Previous research has primarily focused on the development of NLP methods to automatically and effectively detect online hate speech while disregarding further action needed to calm and discourage individuals from using hate speech in the future. In addition, most existing hate speech datasets treat each post as an isolated instance, ignoring the conversational context. In this paper, they propose a novel task of generative hate speech intervention, where the goal is to automatically generate responses to intervene during online conversations that contain hate speech.

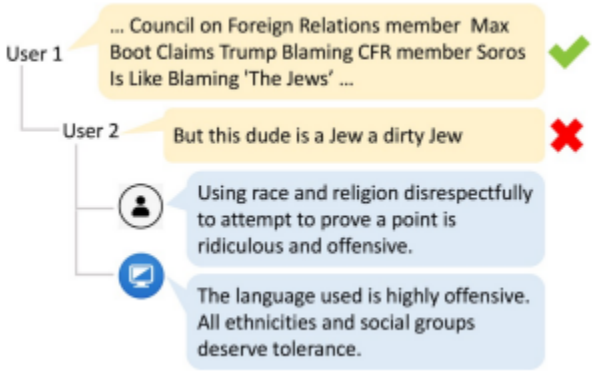

Figure 1: An illustration of hate speech conversation between User 1 and User 2 and the interventions collected for our datasets. The check and the cross icons on the right indicate a normal post and a hateful post. The utterance following the human icon is a *human-written** intervention, while the utterance following the computer icon is machine-generated."

Generalizable hate speech detection can be challenging due to the limitations of existing NLP techniques, dataset construction, and the characteristics of online hate speech, which are frequently mixed up, as mentioned by Zubiaga et al. (2021). Overfitting due to a small dataset and data biases can be carried into models, which is an issue discussed. Hate speech could also come in various forms. These factors add to NLP's generalizability challenge, and researchers suggested that further study on data and model tuning, as well as broader context and effect, open-sourcing, and multilingual research, should be carried out.

Schmidt and Weigan in 2017 discussed the detection of hate speech in the growing body of social media content in the entire world, to determine the rise of hate speech on the platform. This task is usually supervised learning that is commonly done using Natural Language Processing (NLP) techniques and it needs lexical resources

such as the list of slurs, which can help with classification but in order to achieve more accurate results, combination with more features will be the way to go. Linguistic knowledge such as dependency-parse information, and specific linguistic constructs such as imperatives or politeness seem to be effective. In conducting this study, the researchers have also encountered problems such as cultural implications that depend on someone's cultural background. The researchers recommended that for better differentiation of the features, the use of benchmark data sets can be utilized.

In the study of Fortuna et al. (2021), the researchers discussed the use of cross-dataset generalization and possible dataset characteristics and models in generalization in classifying hate speeches. Their study is the first to use a Random Forest classifier to attempt to predict generalization based on dataset attributes and model parameters. They discovered that intra-dataset model performance is the most relevant generalization predictor and identified the kinds of categories that are better suited for training categories for models with generalization potential. It was also found that it is critical to have correct and non-overlapping category definitions. They used a multilingual BERT model, and they observed that the model does not perform well compared to intra-lingual techniques. With that, they still recommend further investigations on how to improve multilingual models by data merging, combining techniques, or modeling,

In the text classification system, the classifier is the main part. The classifier performance quality is directly related to the efficiency and effect of text classification. According to Wei, et al. (2018), most of the classifiers are based on the methods from information retrieval and the machine learning algorithms that are introduced for text classification purposes.

According to Hassan, et al in 2022, Logistic Regression is quick to train data, works well for categorical data, for Simple Parameter Estimation and better for linear data. It is not better for non-linear data and requires a large sample size.

The multilayer perceptron from the study of Alsmadi, et al. (2009) and the recursive neural network from Pouyanfar, et al. (2018) are the first two deep learning approaches used for the text classification task, which improves performance compared with traditional models. CNNs, Recurrent Neural Networks (RNNs), and attention mechanisms are used for text classification. (Qin, L., et al. 2020, Deng, Z., et al. 2021 and Zhao, et al. 2018)

Purpose of the Experiment

American Library Association (2017) defined hate speech as any form of expression that derogates, incites hatred against a group of people or class of persons depending on their race, ethnicity, gender identity, disability or national origin. Hate speech has been around for a long time, not just here in the country but also all around the world. In the age before social media existed, hate speech has also been around and it is already bad, but nowadays, it is a lot more common due to a lot of new avenues in which people communicate.

Social media has been around for a while now, making it easier for people to communicate and share their thoughts without considering the effect of the posts they make. But the rise of social media use has also unlocked a significantly higher amount of expression that revolves around certain hateful rhetoric that the world, at least in the age before computers, has not seen.

Social media algorithms are now capable of showing content or posts related to the user's interest or even exposing them to a wider range of topics to maximize engagement. Given that, it can reach mass audiences with ease and could potentially affect someone's feelings and promote potential crimes, as also mentioned by Laub (2019) in his article. It became an idea for the researchers to come up with an idea to create an experiment to detect hate speech and classify them in order to help in preventing further hate speeches that can potentially harm the person directed to that hate speech. Detecting and classifying it would be an excellent starting point for combating hate speech.

The main purpose of this experiment is to classify hate speeches using Long Short-Term Memory (LSTM) and Softmax Regression. The experiment is inspired by the study of Muslim et al. (2020), where they used it for the Twitter dataset.

According to the research of Almeida F. (2019), word embeddings are highly useful in NLP tasks such as parsing and sentiment analysis, chunking, and question and answer. These advancements are now included in several toolkits, such as Word2Vec, Gensim, and GloVe, allowing for more accurate and less time-consuming usage of word embeddings. With those findings, this experiment will also utilize word embeddings, and the results from LSTM and Softmax Regression will be compared to see which methods will provide better results and be able to classify hate speeches correctly.

Word Embeddings

These are a way to represent words numerically in a d-dimensional vector space.

According to Ameida, F. & Xexeo, G. (2019) the task of representing words and documents is part and parcel of most, if not all, Natural Language Processing (NLP)

tasks. In general, it has been found to be useful to represent them as vectors, which have an appealing, intuitive interpretation, can be the subject of useful operations (e.g. addition, subtraction, distance measures, etc) and lend themselves well to be used in many Machine Learning (ML) algorithms and strategies.

They also state that words with similar contexts (other words) have the same meaning. For example, the words Dostoevsky and Tolstoy may be words that appear in the context of Russian related literature, hypothetically when these words are represented as vectors in a 2-dimensional space they might appear closer to each other.

On a side note we can represent these embeddings in a lower dimensional such that words can be visualized in a 2D plane or 3D plane. T-distributed Stochastic Neighbor Embedding (T-SNE) is a machine learning algorithm for data visualization, which is based on a nonlinear dimensionality reduction technique. The basic idea of t-SNE is to reduce dimensional space keeping relative pairwise distance between points. In other words, the algorithm maps multi-dimensional data to two or more dimensions, where points which were initially far from each other are also located far away, and close points are also converted to close ones. It can be said that t-SNE is looking for a new data representation where the neighborhood relations are preserved.

More recently ways of creating embeddings have surfaced, which rely not on neural networks and embedding layers but on leveraging word-context matrices to arrive at vector representations for words. Among the most influential models is that of the GloVe model by Pennington, J., et al. (2014).

Another model proposed was named ELMo (Embedding from language models) looks at the entire sentence as it assigns each word an embedding. It uses a bi-directional recurrent neural network (RNN) trained on a specific task to create the embeddings. Since it uses a bidirectional architecture, the embedding is based on both the next and previous words in the sentence. Which is useful given that words more often than not have their meaning dependent upon the context they are in or the words surrounding them. Many of the suggested advances seen in the literature have been incorporated in widely used toolkits, such as Word2Vec, gensim, FastText, and GloVe, resulting in ever more accurate and faster word embeddings to be used in NLP.

While various models have been proposed to train embeddings of numerous words such as Word2Vec, GloVe, ELMo etc., sometimes sparsity in data can occur, and deciding to train such word embeddings from scratch may not be wise because it can be difficult with merely a few say thousand words to train a model that represents these words in a vector of high-dimensional space. Fortunately, open source and publicly

pre-trained embeddings can be of great use to an individual looking to use the vector representation of the words they have in their vocabulary or corpus

With the use of these word embeddings particularly GloVe the one we propose to use in this problem we can represent the words in our corpus in a vector space such that it is useful later on in the two models we have chosen to classify and detect certain phrases with derogatory, non-derogatory, offensive, or homonymous words.

In particularly the pre-trained word embeddings we've chosen to use is that of the common crawl pre-trained word vectors also by Pennington, J., et al. (2014), which contains over 42 billion tokens, 1.9 million uncased words in its vocabulary, each with vector representations of 300 dimensions.

Description of the Dataset

The first dataset that will be used for this study will be acquired from the ETHOS Hate Speech Dataset extracted from Github that is used in the study of Qian et al. (2019). The dataset contains 998 comments which include 565 non-hate speech comments and 433 hate speech comments. Those 433 hate speech comments are divided into classes violence and nonviolent hate speech. Distinct labels of those hate comments include gender, race, national origin, disability, and religion.

Besides the first source mentioned, the other two datasets will be coming from the corpus repository of the study entitled “Towards a Comprehensive Taxonomy and Large-Scale Annotated Corpus for Online Slur Usage” by Kurrek J. et al. (2020) and the repository of the paper entitled “Automated Hate Speech Detection and the Problem of Offensive Language” of Davidson T. et al (2019). The 2020 study’s dataset contains 40,000 annotated Reddit comments, in which its features are: id, link_id, parent_id, score, subreddit, author, slur, body, disagreement, and gold_label. The last dataset’s features contain count, hate_speech, offensive_language, neither, and class.

Given the dataset, it was cleaned to remove null, noisy, or unnecessary values. After that, it was integrated into one repository, and data selection was done with the purpose of choosing only the data columns that would be used for the hate speech classifier model. The extracted dataset contains 65780 rows of comments, each with its own id and label. The label values are 3 for homonyms, 2 for derogatory, 1 for non-derogatory, and 0 for offensive comments.

Application of the Experiment (Screenshots)

Exploratory Data Analysis

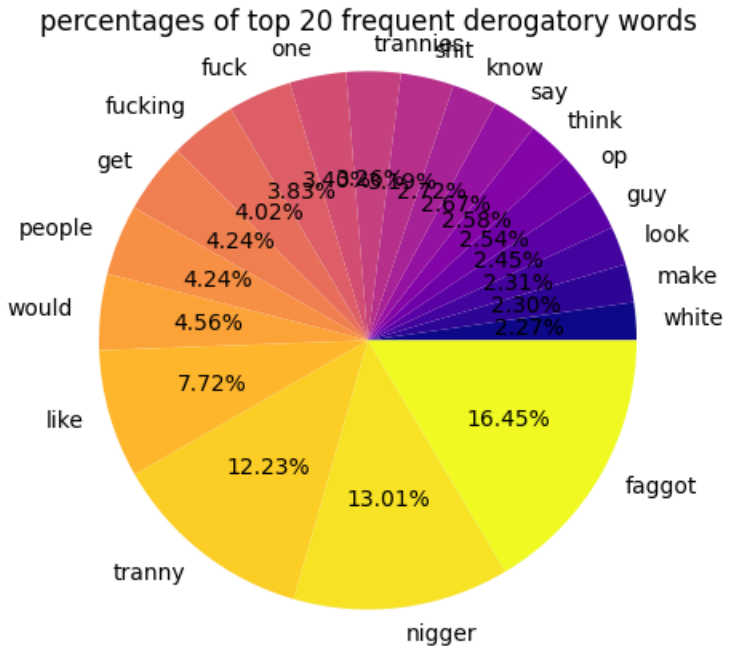

Figure 2. Pie chart of Frequencies/Percentages of Top 20 Derogatory Words

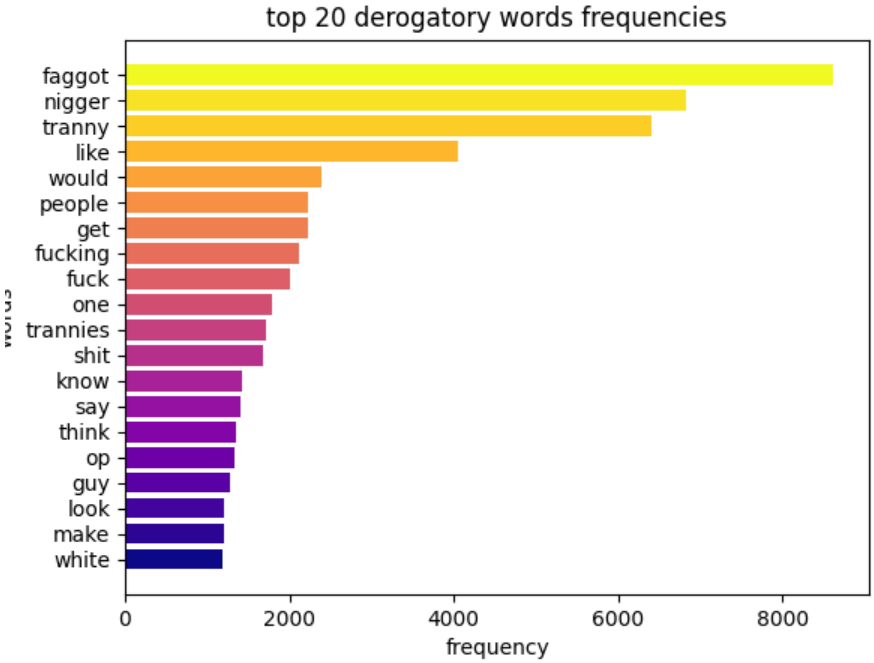

Figure 3. Bar graph of Frequencies of Top 20 Derogatory Words

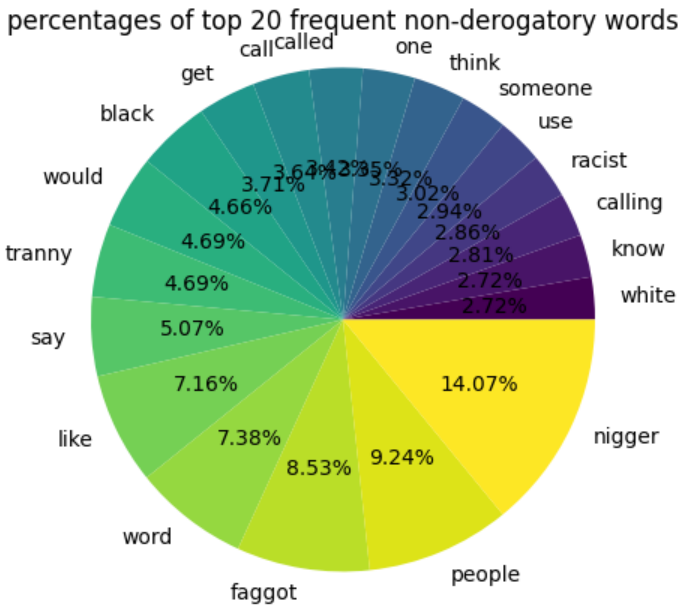

Figure 4. Pie Chart of Frequencies/Percentage of Top 20 Non-Derogatory Words

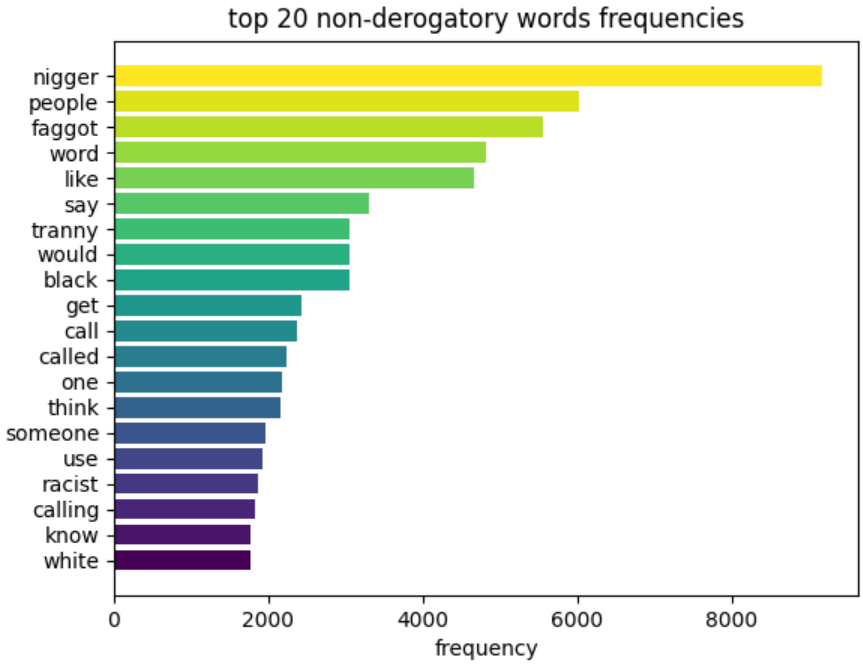

Figure 5. Bar Graph of Frequencies of Top 20 Non-Derogatory Words

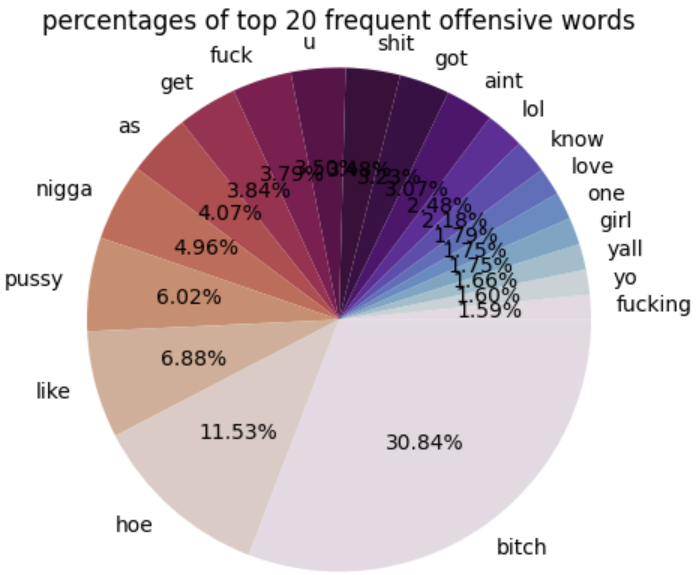

Figure 6. Pie Chart of Frequencies/Percentage of Top 20 Offensive Words

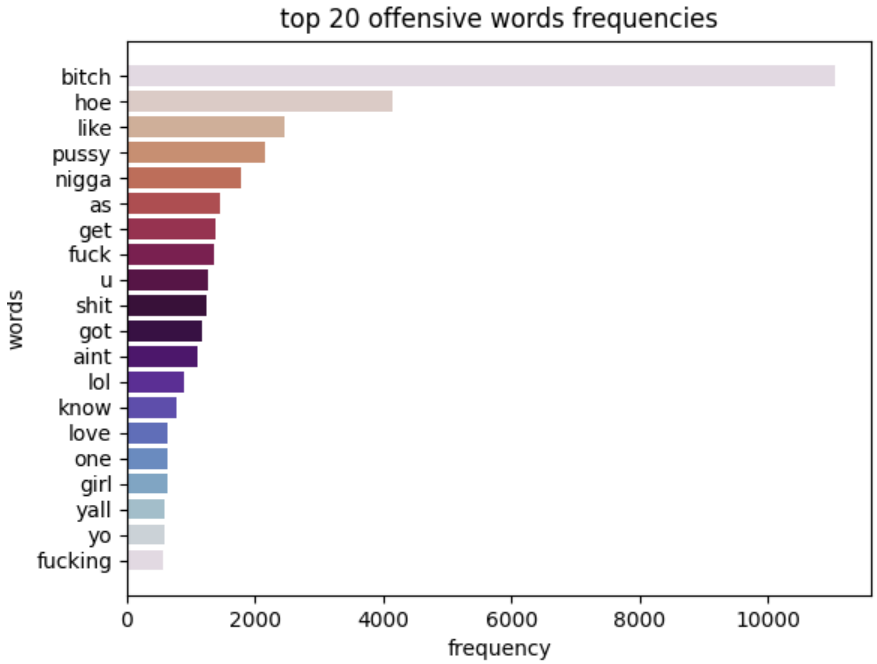

Figure 7. Bar Graph of Frequencies of Top 20 Offensive Words

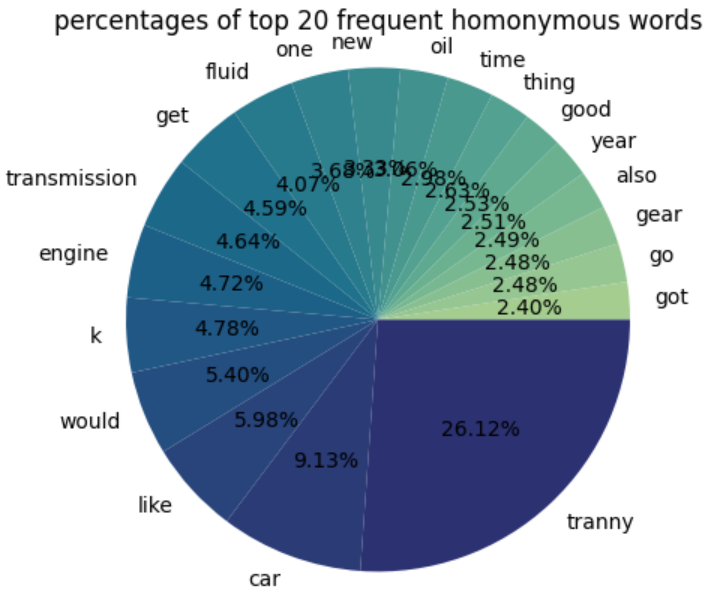

Figure 8. Pie Chart of Frequencies/Percentage of Top 20 Homonymous Words

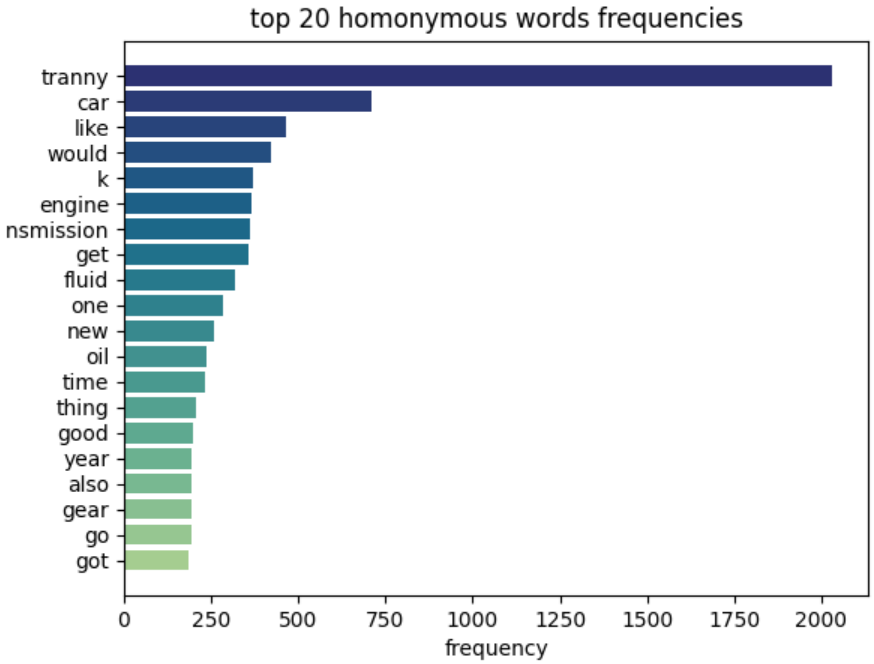

Figure 9. Bar Graph of Frequencies of Top 20 Homonymous Words

Results

LSTM Results

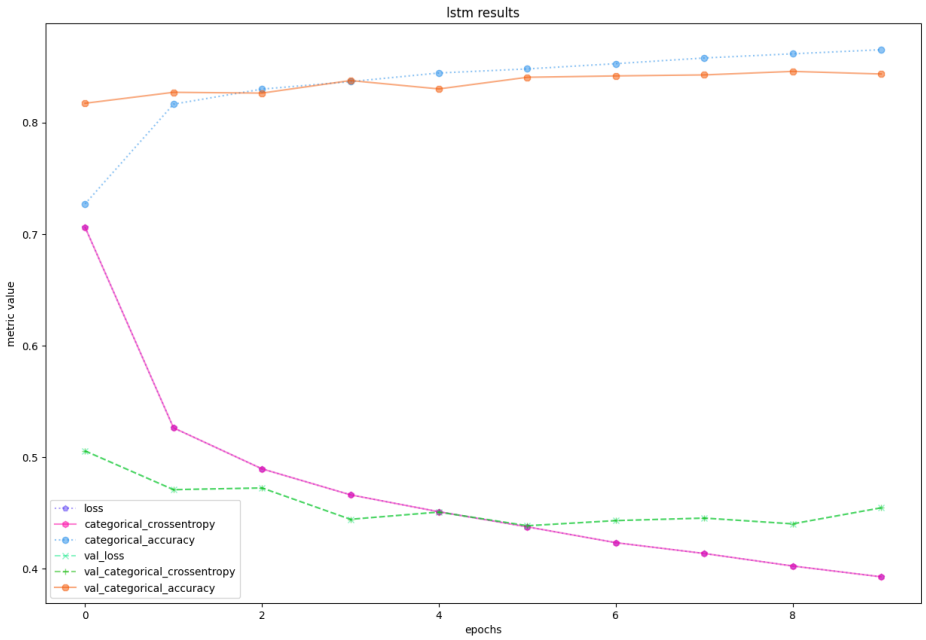

Figure 10. LSTM Classifier Result

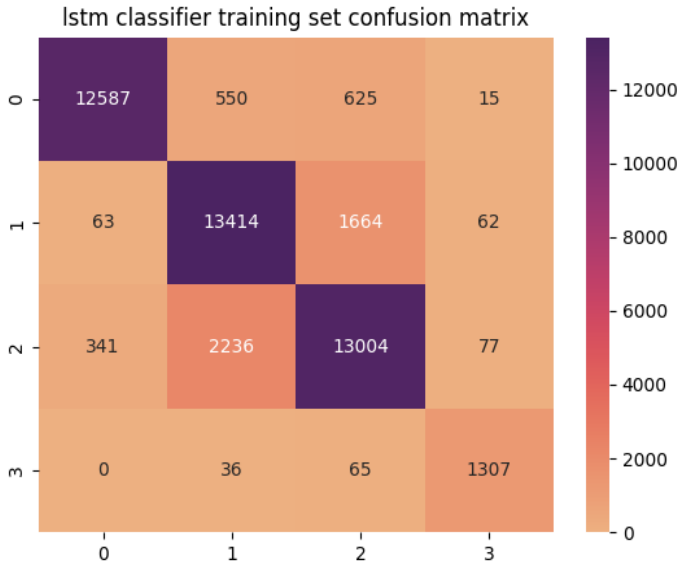

Figure 11. LSTM Training Set Confusion Matrix

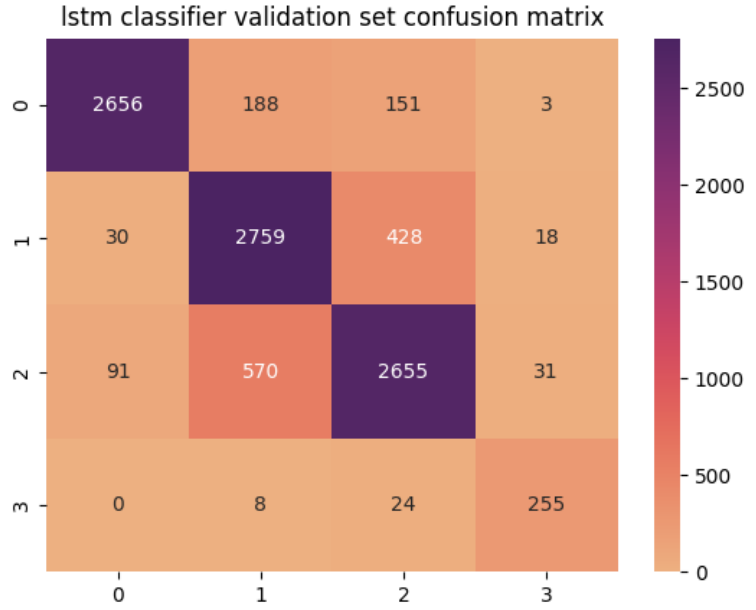

Figure 12. LSTM Validation Set Confusion Matrix

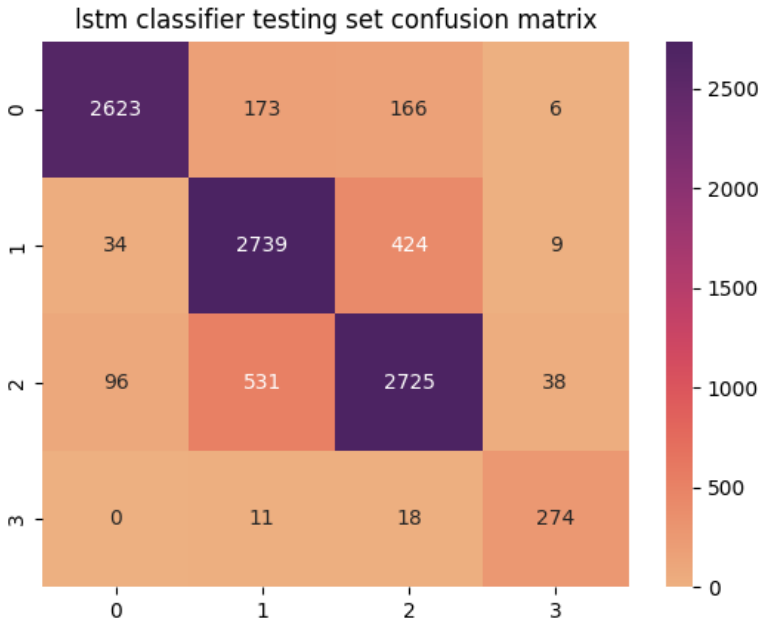

Figure 13. LSTM Testing Set Confusion Matrix

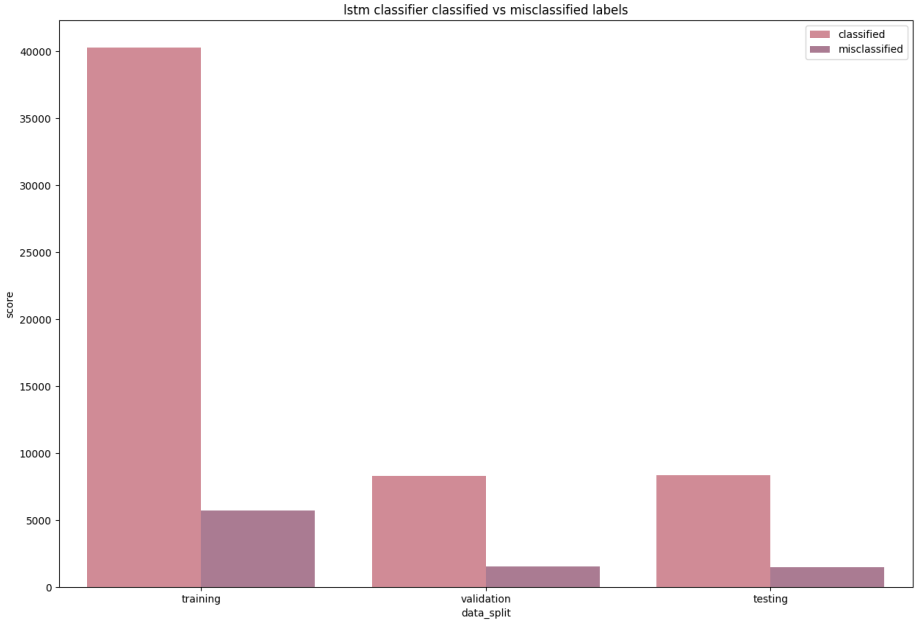

Figure 14. LSTM Classifier Classified vs. Misclassified labels

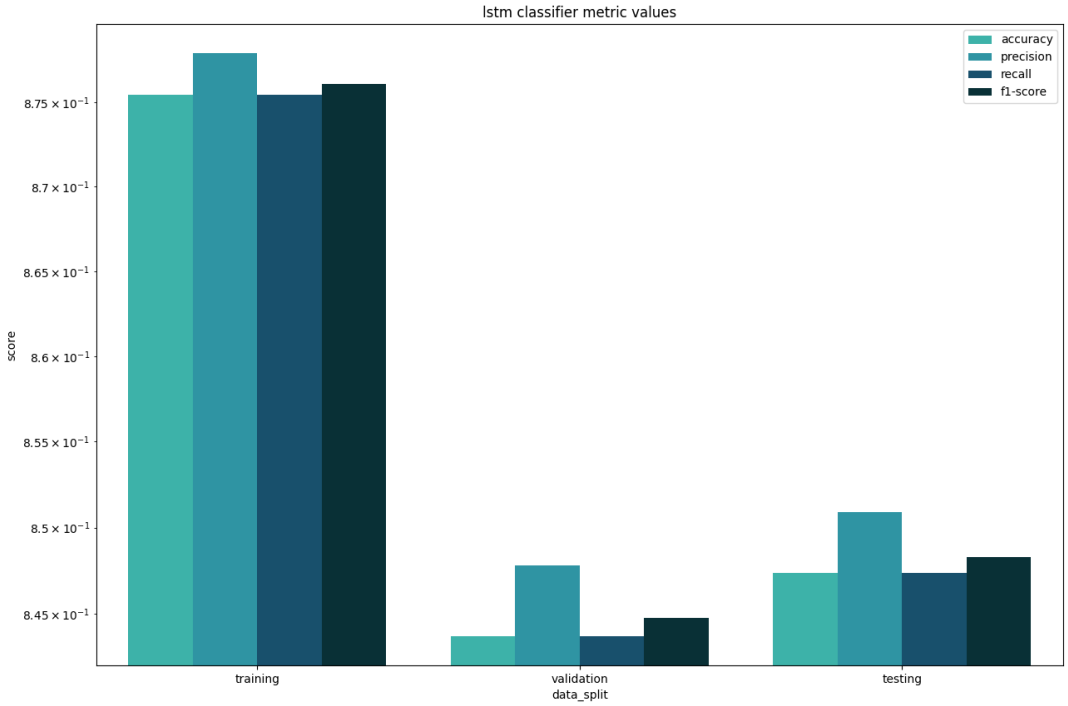

Figure 15. LSTM Metric Values for Training, Validation, and Testing

```
Accuracy: 87.55%
Precision: 87.80%
Recall: 87.55%
F1 score: 87.61%

Accuracy: 84.37%
Precision: 84.78%
Recall: 84.37%
F1 score: 84.48%

Accuracy: 84.74%
Precision: 85.09%
Recall: 84.74%
F1 score: 84.83%
```

Figure 16. Accuracy, Precision, Recall, F1 Score for LSTM Classifier Performance

Softmax Results

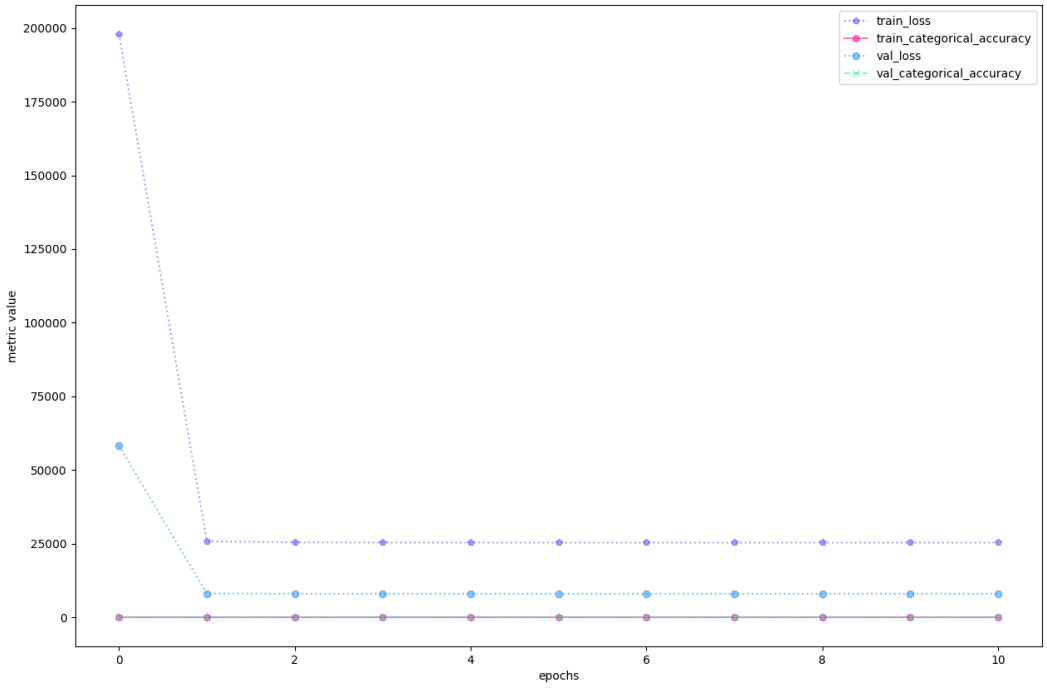

Figure 17. Softmax Classifier Result

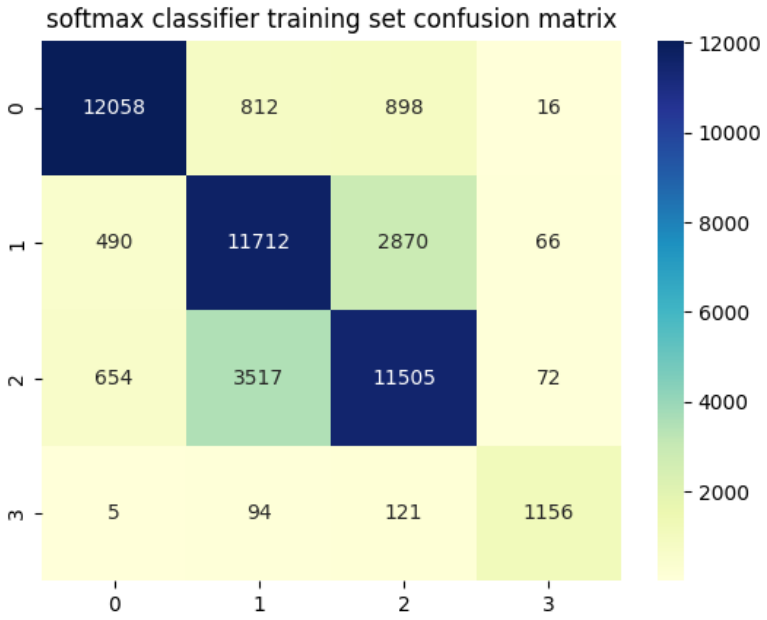

Figure 18. Softmax Classifier Training Set Confusion Matrix

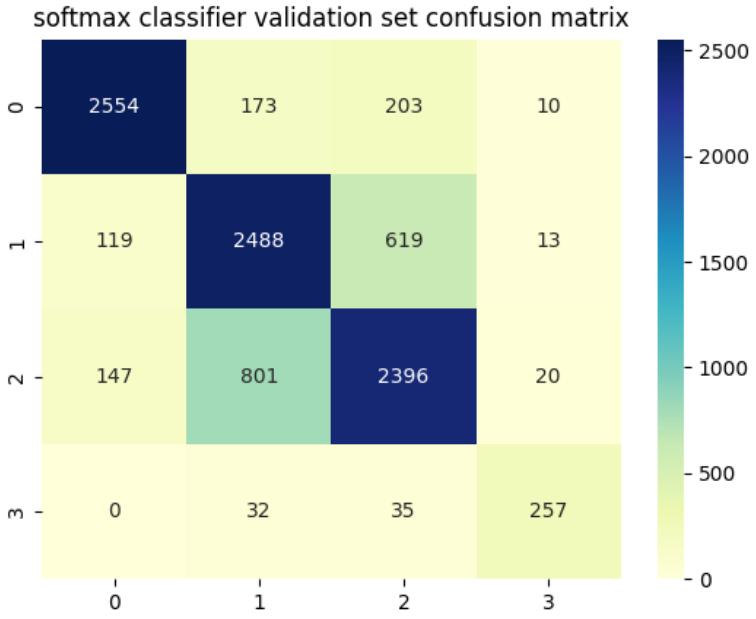

Figure 19. Softmax Classifier Validation Set Confusion Matrix

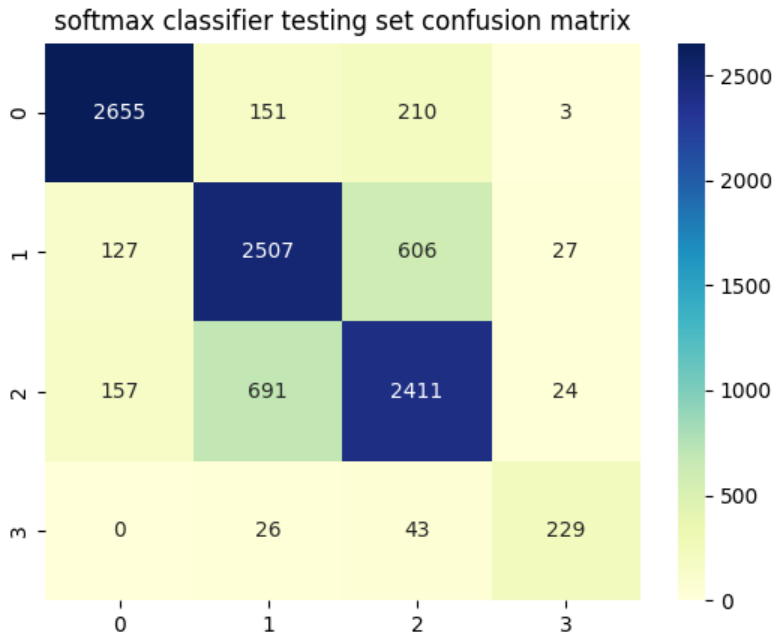

Figure 20. Softmax Classifier Testing Set Confusion Matrix

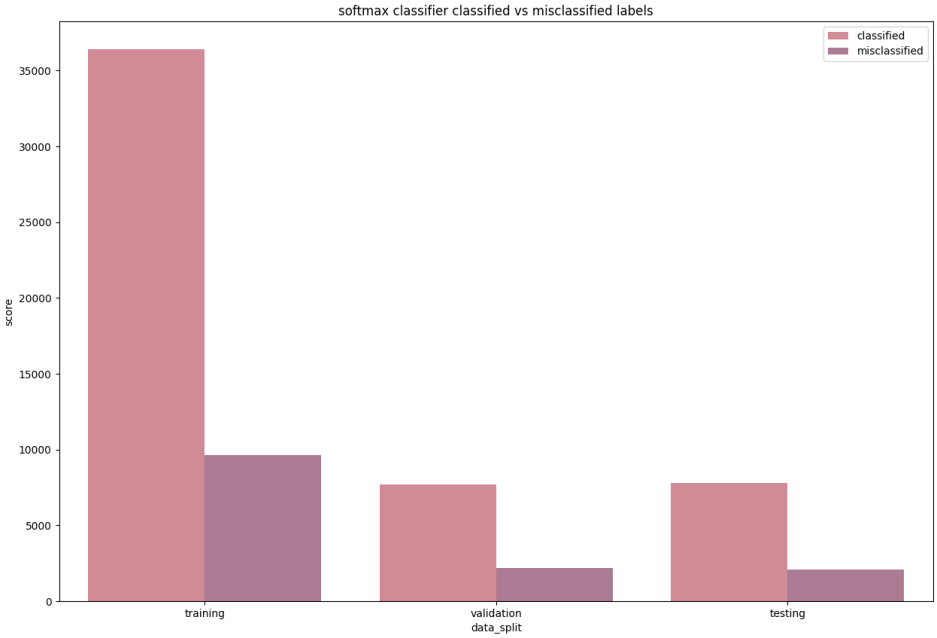

Figure 21. Softmax Classifier Classified vs Misclassified Labels

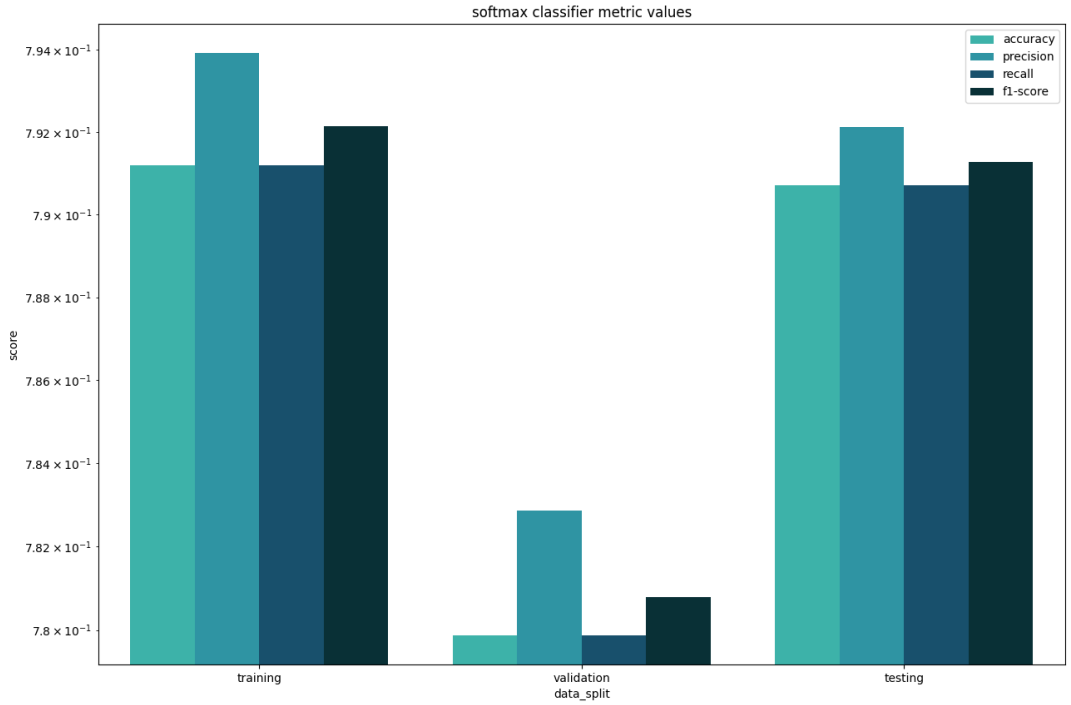

Figure 22. Softmax Classifier Metric Values for Training, Validation, and Testing

```
Accuracy: 79.12%
Precision: 79.39%
Recall: 79.12%
F1 score: 79.21%

Accuracy: 77.99%
Precision: 78.29%
Recall: 77.99%
F1 score: 78.08%

Accuracy: 79.07%
Precision: 79.21%
Recall: 79.07%
F1 score: 79.13%
```

Figure 23. Accuracy, Precision, Recall, F1 Score for Softmax Classifier Performance

Summary of Findings

The created program has been able to identify into four categories: derogatory, non-derogatory, offensive, and homonymous words. The program was also able to determine the number of words and the percentage of each word.

The top five derogatory words that are classified are: faggot with 16.45%, nigger with 13.01%, tranny with 12.23%, fucking with 4.02%, and fuck with 3.83%. While the top five non-derogatory words that are classified are: nigger with 14.07% people with 9.24%, faggot with 8.53%, word with 7.38%, and like with 7.16%. Then, the top five offensive words are: bitch with 30.84%, hoe with 11.53%, pussy with 6.02%, nigga with 4.96%, and fuck with 3.79%. Lastly, the top five homonymous words are: tranny with 26.12%, car with 9.13%, like with 5.98%, would with 5.40%, and k with 4.78%.

The LSTM Classifier performed better in classifying hate speeches with an f-score of 87.61% for the training set, 84.48% for the validation set, and 84.83% for the testing set. It accurately classified labels 40,312 times during training, 8325 times during validation, and 8361 times during testing. It also misclassified labels 5734 times during training, 1542 times during validation, and 1506 times during testing.

The Softmax Classifier performed worse with an f-score of 79.21% for the training set, 78.08% for the validation set, and 79.13% for the testing set. It accurately classified labels 36431 times during training, 7695 times during validation, and 7802 times during testing. It also misclassified labels 9615 times during training, 2172 times during validation, and 2065 times during testing.

Conclusion

Based on the findings of the study, the following conclusions were drawn:

1. The LSTM has outperformed the Softmax Regression in classifying hate speeches. The LSTM has an accuracy of 87.55%, precision of 87.80%, recall of 87.55%, and an f-score of 87.61% for the training set, the validation set has an accuracy of 84.37%, precision of 84.78%, recall of 84.37% and an f-score of 84.48%, while the testing set has an accuracy of 84.74%, precision of 85.09%, recall of 84.74% and an f-score of 84.83%.
2. The Softmax Regression on the other hand has performed worse with an accuracy of 79.12%, precision of 79.39%, recall of 79.12%, and an f-score of 79.21% for the training set, the validation set has an accuracy of 77.99%, precision of 78.29%, recall of 77.99% and a f-score of 78.08%, while the testing set has an accuracy of 79.07%, precision of 79.21%, recall of 79.07% and a f-score of 79.13%.
3. Based on the gathered data from the experiment, it shows that LSTM has performed better in classifying hate speeches compared to the Softmax

Regression approach. This shows that there is a noticeable difference between the results of the two classification models.

References

Almeida, F., & Xexeo, G. (2019). *(PDF) word embeddings: A survey - researchgate*. Word Embeddings: A Survey. https://www.researchgate.net/publication/330700931_Word_Embeddings_A_Survey

Alsmadi, M.K., Omar, K.B., Noah, S.A., & Almarashdah, I. (2009). Performance Comparison of Multi-layer Perceptron (Back Propagation, Delta Rule and Perceptron) algorithms in Neural Networks. *2009 IEEE International Advance Computing Conference*, 296-299.

American Library Association. (2017). "Hate Speech and Hate Crime". http://www.ala.org/advocacy/intfreedom/hate (Accessed May 22, 2023)

Deng, Z., Peng, H., He, D., Li, J., & Yu, P. S. (2021). HTCInfoMax: A global model for hierarchical text classification via information maximization. *arXiv preprint arXiv:2104.05220*.

Fortuna, P., Soler-Company, J., & Wanner, L. (2021, February 9). *How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets?*. Information Processing & Management. https://www.sciencedirect.com/science/article/pii/S0306457321000339?ref=pdf_download&fr=RR-2&rr=7dfefa11186dbc46

Hassan, S.U., Ahamed, J., Ahmad, K. (2022). Analytics of machine learning-based algorithms for text classification. Sustainable Operations and Computers. 3. 238-248, ISSN 2666-4127. https://doi.org/10.1016/j.susoc.2022.03.001.

Kurrek, J., Saleem, H. M.; Ruths, D. (n.d.). Towards a comprehensive taxonomy and large-scale annotated corpus for online slur usage. ACL Anthology. https://www.aclweb.org/anthology/2020.alw-1.17

Laub, Z. (2019). *Hate speech on Social Media: Global Comparisons*. Council on Foreign Relations. https://www.cfr.org/backgrounder/hate-speech-social-media-global-comparisons#chapter-title-0-3

Li, Q., Peng, H., Li, J., Xia, C., Yang, R., Sun, L., Yu, P.S. and He, L.. (2021). A Survey on Text Classification: From Traditional to Deep Learning. ACM Trans. Intell. Syst. Technol. 37, 4, Article 111 (April 2021), 39. https://doi.org/10.1145/1122445.1122456

Mollas, I., Chrysopoulou, Z., Karlos, S. *et al. (2022).* ETHOS: a multi-label hate speech detection dataset. *Complex Intell. Syst.* 8, 4663–4678. https://doi.org/10.1007/s40747-021-00608-2

Muslim, A., Mutiara, A. B., Refianti, R., Karyati, C. M., & Setiawan, G. (2020). *Comparison of accuracy between long short-term memory-deep learning and multinomial logistic regression-machine learning in sentiment analysis on Twitter*. International Journal of Advanced Computer Science and Applications (IJACSA). https://thesai.org/Publications/ViewPaper?Volume=11&Issue=2&Code=IJACSA&SerialNo=94

Qian, J., Bethke, A., Liu, Y., Belding, E., Yang Wang, W. (2019). A Benchmark Dataset for Learning to Intervene in Online Hate Speech. https://doi.org/10.48550/arXiv.1909.04251

Qin, L., Che, W., Li, Y., Ni, M., & Liu, T. (2020, April). Dcr-net: A deep co-interactive relation network for joint dialog act recognition and sentiment classification. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 34, No. 05, pp. 8665-8672).

Pouyanfar, S., Sadiq, S., Yan, Y., Tian, H., Tao, Y., Reyes, M. P., ... & Iyengar, S. S. (2018). A survey on deep learning: Algorithms, techniques, and applications. *ACM Computing Surveys (CSUR)*, *51*(5), 1-36.

Schmidt, A. and Wiegand, M. (2017). A Survey on Hate Speech Detection using Natural Language Processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics. https://aclanthology.org/W17-1101/

T-Davidson. (n.d.). T-Davidson/hate-speech-and-offensive-language: Repository for the paper “Automated hate speech detection and the problem of offensive language”, ICWSM 2017. GitHub. https://github.com/t-davidson/hate-speech-and-offensive-language

Wei, L, Wei, B. and Wang, B. (2012). Text Classification Using Support Vector Machine with Mixture of Kernel. Journal of Software Engineering and Applications. 05. 55-58. 10.4236/jsea.2012.512B012.

Yin, W., Zubiaga, A., S, A., A, A., A, A.-H., H, A.-D., HS, A., AM, A., KM, M., W, A., P, B., H, L., ML, W., A, A., J, P., B, P., P, B., M, G., V, V., … K, M. (2021, June 17). *Towards generalisable hate speech detection: A review on obstacles and solutions*. PeerJ

Computer Science. https://peerj.com/articles/cs-598/?ref=https%3A%2F%2Fgithubhelp.com

Zhao, W., Ye, J., Yang, M., Lei, Z., Zhang, S., & Zhao, Z. (2018). Investigating capsule networks with dynamic routing for text classification. *arXiv preprint arXiv:1804.00538*.