iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** VII **Month of publication:** July 2022

DOI: https://doi.org/10.22214/ijraset.2022.46099

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VII July 2022- Available at www.ijraset.com

Intricate Twitternlp Modelling

Aditi Ashish Gawande[1], S Karthikeyan[2], Sriram Balasubramanian[3], S Ajay[4]
[1, 2, 3, 4]*CS Department, SRM University KTR University*

Abstract: ***In this research, we introduce TweetNLP, a platform for social media Natural Language Processing (NLP). An extensive range of NLP tasks are supported by TweetNLP, including standard focus areas like sentiment analysis and named entity recognition as well as social media-specific tasks like emoticon prediction and offensive language detection. Task-specific systems run on moderately small Transformer-based language models that are focused on social media text, particularly Twitter, and don't require specialized hardware or cloud services to operate. TweetNLP's major contributions are: (1) an integrated Python library for a contemporary toolkit supporting social media analysis using various task-specific models tailored to the social domain; (2) an interactive online demo for codeless experimentation using our models; and (3) a tutorial covering a wide range of typical social media applications.***
Keywords: ***NLP, Sentiment Analysis, Emotion Recognition, Emoji Prediction, Hate Speech Detection***

I. INTRODUCTION

Social media is characterised by its connectivity, accessibility, and content creation. It is a useful tool for sharing, creating, and disseminating information as well as for communicating with people locally and globally. Social media usage has evolved into a regular activity in today's world. Twitter, Instagram, Youtube and other social media platforms have all emerged as major informational resources.

It has been discovered that, by extracting and analysing data from social networking sites, an understanding of contemporary society can be developed. Online users communicate with each other by sending text-only messages or enhancing them with multimedia content like images, audio, or video. This has led to the usage of these platforms to comprehend user, group, and organisational behaviour. Particularly, twitter, the primary medium examined in this work, has long been a valuable tool for comprehending society as a whole. Twitter is a crucial research and practical resource for natural language processing (nlp) because of its relevance and accessibility.

Twitter is intriguing for nlp because it embodies many characteristics that come naturally in fast-paced, impromptu conversation. The improvement of results on benchmark datasets with roughly independent and identically distributed (iid) training, validation, and testing sections, drawn from data that was gathered or validated by open sourcing, has been the focus of a significant and influential thread of research on natural language understanding (nlu).

Additionally there are significant flaws in allegedly high-performing systems, and they nonetheless lack human-level task competence. In fact, it has been demonstrated that even conventional nlp systems perform poorly when applied to social media, particularly when performing tasks like normalisation, part-of-speech tagging, sentiment analysis, or named entity recognition because of problems like noise, length restrictions for messages related to platforms, jargon, emoticon, colloquial language and multilinguality.

Tweetnlp (tweetnlp.org) provides a library tailored to twitter. Transformer-based language models that have been trained on twitter make up the core of tweetnlp (barbieri et al., 2020, 2022; loureiro et al., 2022). These specialised language models have then undergone additional fine-tuning for particular nlp tasks on twitter data.

All of these resources are consolidated into one platform by tweetnlp. Tweetnlp provides a simple python api that makes it simple to use social media models.

Despite the tendency toward progressively larger language models (shoeybi et al., 2019; brown et al., 2020), tweetnlp is more concerned with the general user and applicability and hence include base models that are simple to operate on standard computers or on free cloud services. The ability to test models and conduct real-time analysis on twitter is provided by an interactive online demo that provides access to all models.

©IJRASET: All Rights are Reserved

S.N.o.	Paper Title	Year & Journal	Description and findings	Inference
1.	BERTweet: A pre-trained language model for English Tweets	2020 Association for Computational Linguistics	Same architecture as BERTbase, which is trained with a masked language modeling objective. BERTweet pre-training procedure is based on RoBERTa which optimizes the BERT pre-training approach for more robust performance. The model is optimized using Adam (Kingma and Ba, 2014), and uses a batch size of 7K across 8 V100 GPUs (32GB each) and a peak learning rate of 0.0004. BERTweet is pre-trained for 40 epochs in about 4 weeks.	BERTweet outperforms strong baselines RoBERTabase and XLM-Rbase (Conneau et al., 2020), producing better performance results than the previous state-of-the-art models on three Tweet NLP tasks: Part-of-speech tagging, Named-entity recognition and text classification.
2.	RoBERTa: A Robustly Optimized BERT Pretraining Approach	2019		Performance can be substantially improved by training the model longer, with bigger batches over more data; removing the next sentence prediction objective; training on longer sequences; and dynamically changing the masking pattern applied to the training data. RoBERTa achieves state-of-the-art results on GLUE, RACE, and SQuAD, without multi-task fine-tuning for GLUE or additional data for SQuAD.
3.	TimeLMs: Diachronic Language Models from Twitter	2022 Association for Computational Linguistics	A variety of qualitative evaluations to demonstrate how they respond to patterns and peaks in an activity involving certain named things or idea drift. Lack of diachronic specialization is especially concerning in contexts such as social media, where topics of discussion change often and rapidly. We address this issue by sharing with the community a series of time-specific LMs specialized in Twitter data.	A quantitative analysis on the degradation suffered by language models over time; the relation between time and size; a qualitative analysis where they show the influence of time in language models for specific examples.
4.	T-NER: An All-Round Python Library for Transformer-based Named	2021 Association for Computational Linguistics	T-NER facilitates the study and investigation of the cross-domain and cross-lingual generalization ability of LMs fine-tuned on NER. In-domain performance is generally competitive across datasets. However, cross-domain	This paper especially focuses on LM finetuning, and empirically shows the difficulty of cross-domain generalization in NER.

	Entity Recognition		generalization is challenging even with a large pre-trained LM, which has nevertheless capacity to learn domain-specific features if finetuned on a combined dataset.	They have also facilitated the evaluation by unifying some of the most popular NER datasets in the literature, including languages other than English. Which will definitely emphasize the importance of NER generalization analysis.
5.	XLM-T: Multilingual Language Models in Twitter for Sentiment Analysis and Beyond	2022	Multilingual LMs integrate streams of multilingual textual data without being tied to one single task, learning general-purpose multilingual representations. This is an important consideration, as there is ample agreement that the quality of LM-based multilingual representations is strongly correlated with typological similarity. results suggest that when fine-tuning task-specific Twitter-based multilingual LMs, a domain-specific model proves more consistent than its general domain counterpart, and that in some cases a smart selection of training data may be preferred over largescale fine-tuning on many languages.	This paper bridges this typological similarity gap by introducing a toolkit for evaluating multilingual Twitter-specific Language Models. It comprises a large multilingual Twitter-specific LM based on XLMR checkpoints A unified dataset is devised in 8 languages for sentiment analysis (which we call Unified Multilingual Sentiment Analysis Benchmark, UMSAB)

II. SUPPORTING TASKS AND EMBEDDINGS

Discussing the tasks supported by TweetNLP.For classification tasks, we simply fine-tune the models which are described in the TweetEval library, and for refining named entity recognition, we depend on the T-NER library, which is also integrated into TweetNLP.

1) *Sentiment Analysis*- Sentiment analysis is the process of identifying and categorizing the emotions represented in a text source. When analyzed, tweets may produce a significant quantity of sentiment data. These statistics help us understand how individuals feel about a range of issues. Aims to forecast the feeling of a tweet that has one of the three classifications: positive, negative, or neutral.
2) *Emotion Recognition*- Emotions are considered of utmost importance as they have a key responsibility in human interaction. The goal of this activity is to match the most relevant emotion to a tweet, boiled down to surprise, love, hate, boredom, anger, happiness, sadness, and empty.
3) *Emoji Prediction*- The goal of emoji prediction is to predict the final emoji of a tweet, including 20 emoji labels.It is important to take into account various selection heuristics, such as choosing the first or most prevalent emoji inside a tweet, while attempting to identify the genuine label from a variety of emojis in a tweet.
4) *Hate Speech Detection*- The hate speech dataset consists of identifying tweets that are hostile toward immigrants or women. The categories of the dataset in which this model was fine-tuned are hate speech, offensive but not hate speech, or neither offensive nor hate speech.
5) *Offensive Language Identification*- It is crucial to research the detection of abusive language on social media to prevent conflicts brought on by the usage of such language and to make social media platforms safer for kids and teenagers.The assignment is to find any derogatory words in a tweet.

A. *Embeddings*

1) *Word Embeddings*- Social media has the actual ability to injure individuals and may serve as a platform for the spread of racism, misogyny, and other hateful ideologies. Although having the right to free expression is crucial, there are instances when it's necessary to spot such stuff and stop it from being published. In the NLP approach known as word embedding, each word is represented as an n-dimensional vector that depicts the word's projection in vector space. A particular word's position in the model is determined during training by the words that surround it. Word2Vec and GloVe are the two most popular techniques used to build this kind of model.
2) *Tweet Embeddings*- The syntactic and semantic components of a word are both captured by its embedding. Tweets vary from other forms of text in that they are brief, loud, and have particular lexical and semantic characteristics. Word embeddings must be specially learnt from tweets as a result. We selected one response at random for tweets that received several replies. Each mini-batch in training consists of a list of tweet-reply pairings. The enumeration of all other conceivable combinations of tweet-reply, tweet-tweet, and tweet-reply pairings is considered a negative sample; the tweet-reply pairs are considered positive samples. In this paper, we also present experiments demonstrating how to use the data sets in some NLP tasks, such as tweet sentiment analysis and hate recognition, emotion recognition, emoji prediction, and offensive language detection.

III. METHODOLOGY

A. *Sentiment Analysis*

The data input is a test and train dataset containing various tweets and comments and the tweets are of mixed sentiments, such as positive, negative, and neutral. The distribution of Training and Testing data is depicted through a histogram using visualization tools such as Seaborn and Matplotlib.

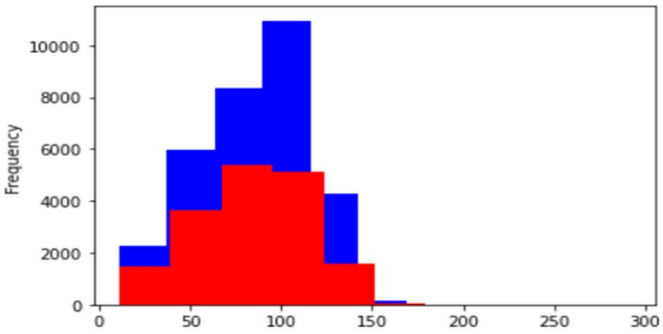


Fig1- The Frequency of Words

The training data is used to train the model in order for it to understand the different words and related contextual sentiment for further analysis with the test data. According to the analysis based on the training data, the most repetitive words within the dataset are the following. By the use of a few different packages within the python environment, we can find out the vocabulary and different sentiment-related words.

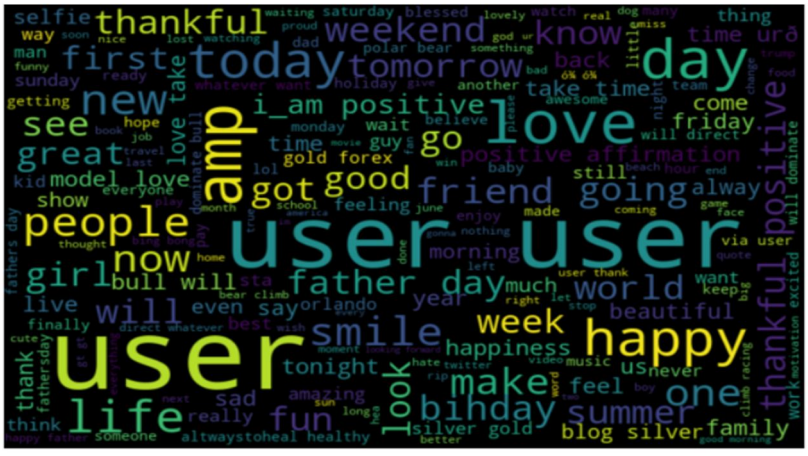


Fig2- The Neutral Words

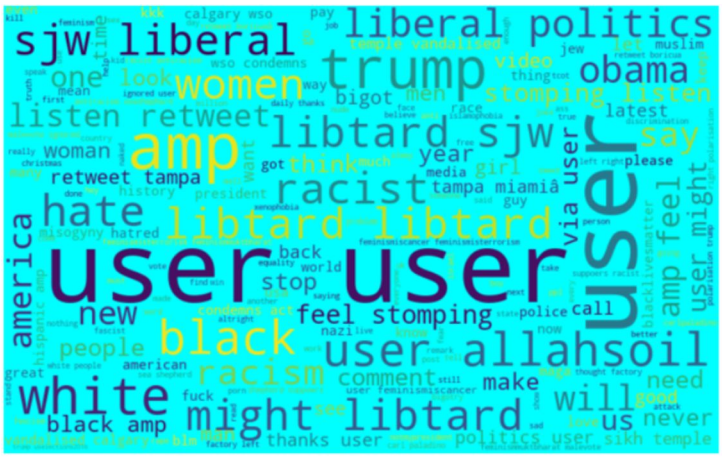


Fig3- The Negative Words

The extraction of the number of tweets that contain any desired hashtags is implemented.
The main reason behind the requirement of hashtag analysis is to find negative comments such as racist tweets, abusive comments, etc.

Analysis of hashtags with regular and neutral tweets, the top 20 most frequently used hashtags are:

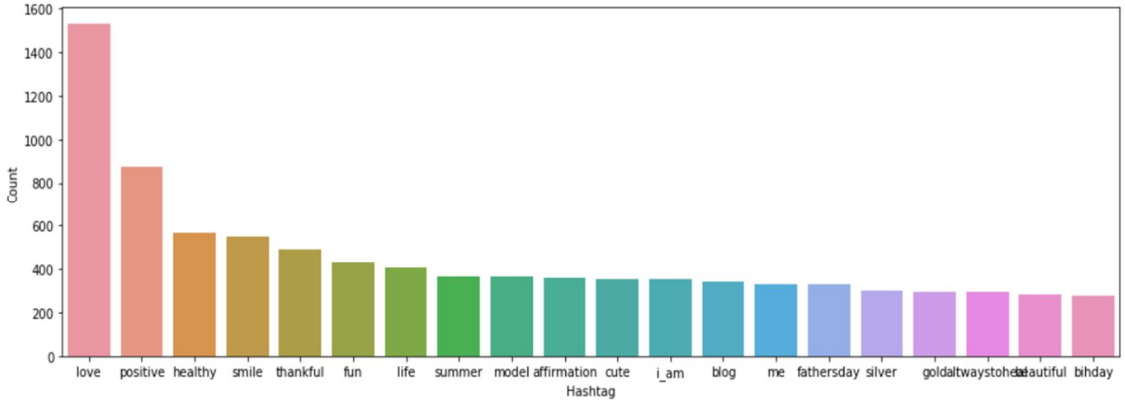


Fig4.1- Top 20 Frequently used regular hashtags

Analysis of hashtags with negative tweets, the top 20 most frequently used hashtags are:

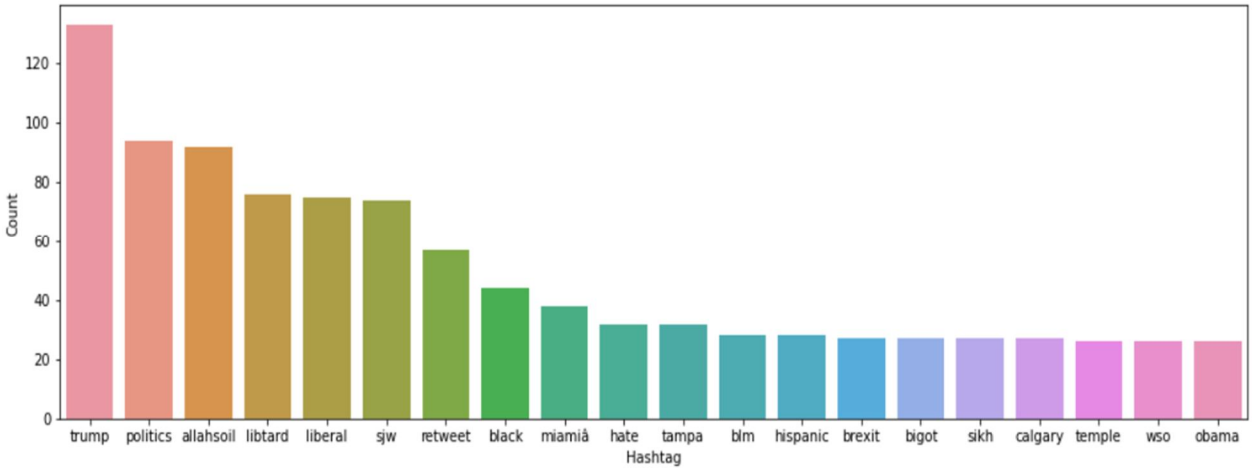


Fig4.2- Top 20 Frequently used negative hashtags

©IJRASET: All Rights are Reserved

B. Emoji Prediction

A dataset with different tweets are kept as the baseline for training the model to predict the correct and relevant emoji at the end of every tweet.

	Tweet	Label
18773	I am so proud of you haley_pepper !!! Love you...	9
19190	#friendshipday #pooh #magickingdom #familyvaca...	5
44506	This #beauty beaded dangle cuff #bracelet has ...	2
10350	Amazing!! #Repost @user Halloween everyone! In...	1
19909	Flashback Friday to this night @ Little Rock, ...	4

Fig5- Loaded Data

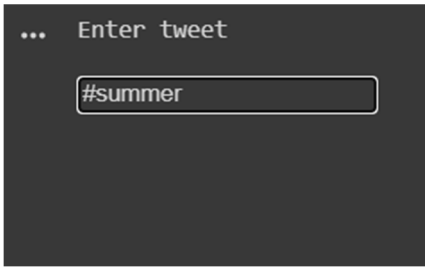
A set of 20 emoji labels with their meanings are mapped to train the model to predict the emoji accurately according to the tweet.

0	❤	Red heart
1	😍	Smiling face with hearteyes
2	😂	Face with tears of joy
3	💕	Two hearts
4	🔥	Fire
5	😊	Smiling face with smiling eyes
6	😎	Smiling face with sunglasses
7	✨	Sparkles
8	💙	Blue heart
9	😘	Face blowing a kiss
10	📷	Camera
11	US	United States
12	☀	Sun
13	💜	Purple heart
14	😉	Winking face
15	💯	Hundred points
16	😁	Beaming face with smiling eyes
17	🎄	Christmas tree
18	📸	Camera with flash
19	😜	Winking face with tongue

Fig6- Mapped Emojis

The true available emoji and the predicted emoji are compared after the training of the model.

The user now can simply enter the desired tweet and get the output as the same tweet with the appropriate emoji following it.



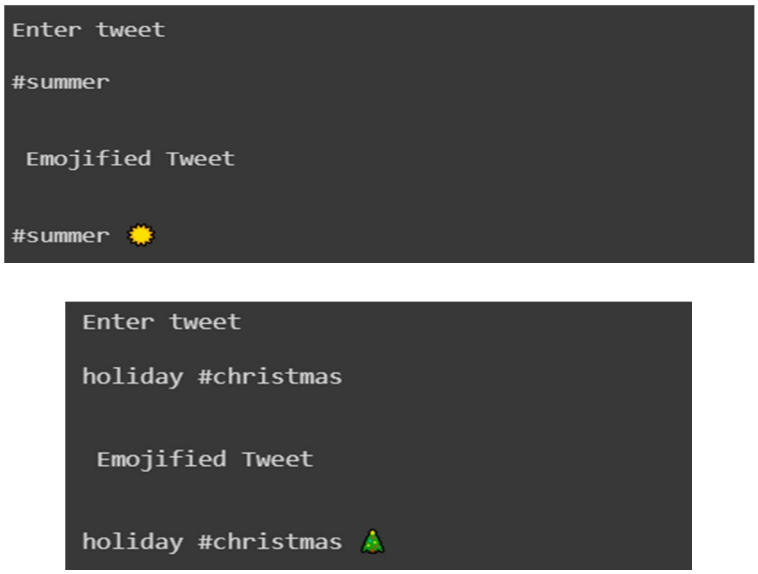


Fig7- Screenshots of the Emoji Prediction Demo

C. Emotion Recognition

The ability to recognize emotions has several uses, including the ability to identify psychological problems like anxiety or sadness in people or gauge how a community feels about a certain issue. In human-computer interaction systems and their applications, emotion recognition is essential. These days, the majority of individuals use social media sites like Twitter, Facebook, Instagram, and others extensively to express their feelings or opinions about a certain subject. As a result, these sites serve as enormous data warehouses for emotional information.

The following is the graphical representation of emotion that we classified from the dataset we used

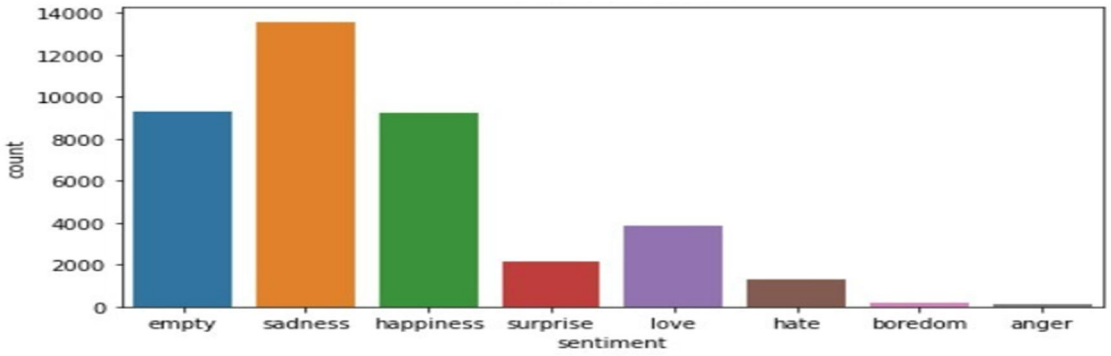


Fig8- Data Distribution

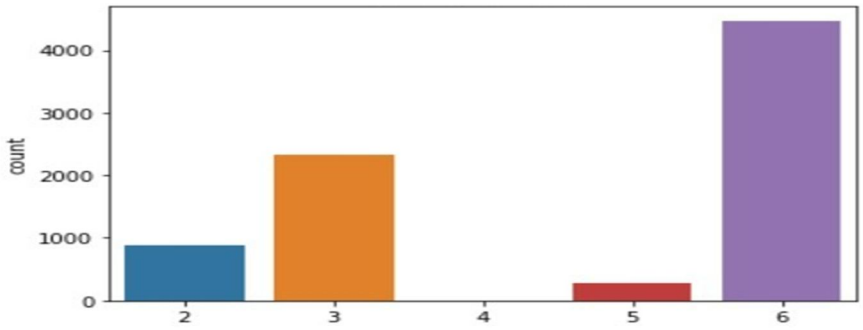


Fig9- Predicted Labels

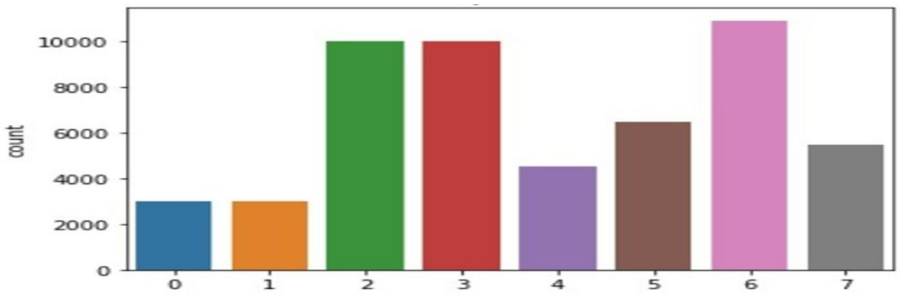


Fig10- Over Sampled Train Data

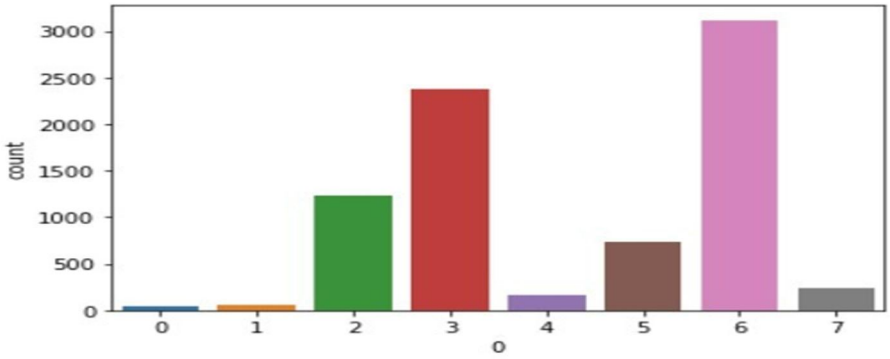


Fig11- Predicted Labels

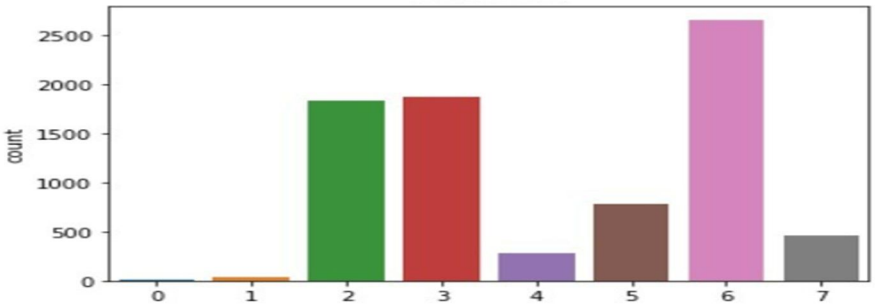


Fig12- Test Labels

D. Hate Speech Detection

The dataset used is from a study called Automated Hate Speech Detection and the Problem of Offensive Language in which 6% of the tweets were categorized as hate speech. The labels on this dataset were voted on by crowdsource and determined by majority rules.

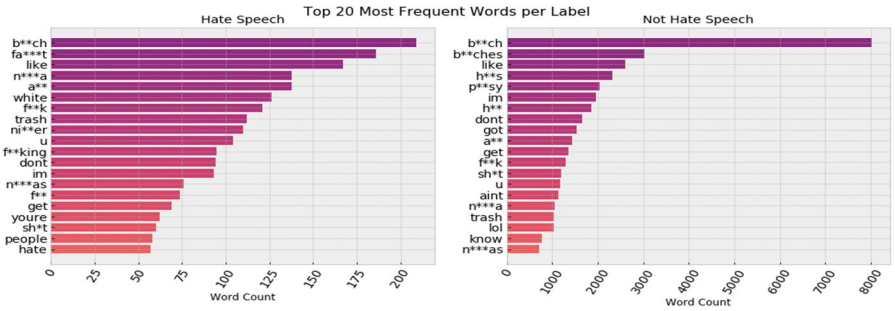


Fig13- Top 20 Most Frequent Words per Label

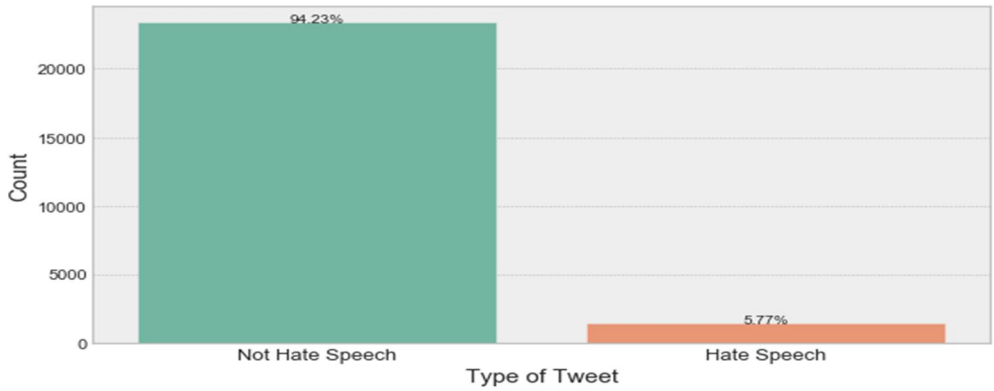


Fig14- Amount of Tweets per Label

A condensed representation used in information retrieval and natural language processing is called the bag-of-words method. In this method, syntax and even word order are ignored while maintaining multiplicity, and a text such as a phrase or document is represented as the bag (multiset) of its words.

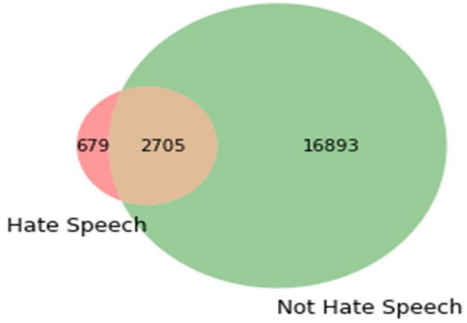
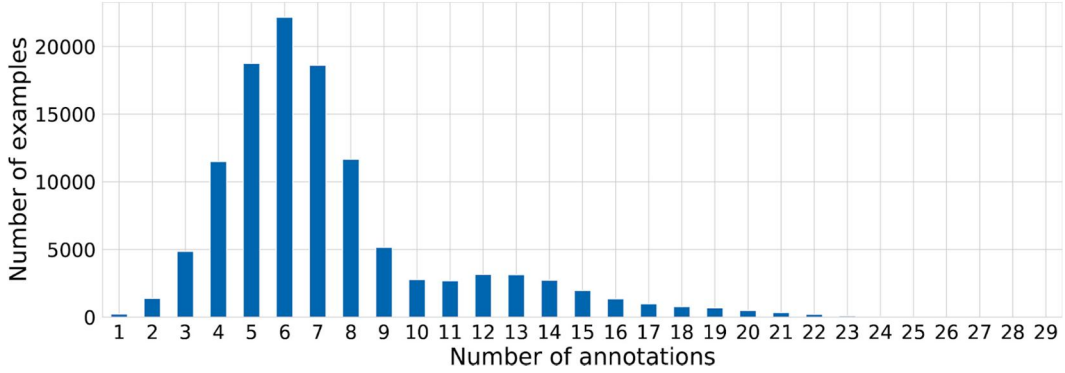


Fig15- Comparison of Unique words in Each Label

E. Offensive Language Identification

Social networks have been increasingly popular in recent years. The idea behind social media was to allow us to express our opinions online, stay in touch with loved ones, and share happy moments. However, as reality is not so ideal, there are others who share hate speech-related messages, use it to abuse particular people, for example, or even build robots whose sole purpose is to attack particular circumstances or individuals. It is difficult to determine who created such content, but there are numerous approaches that might be used, such as natural language processing or machine learning algorithms that can look into the text and make predictions using the related meta-data.



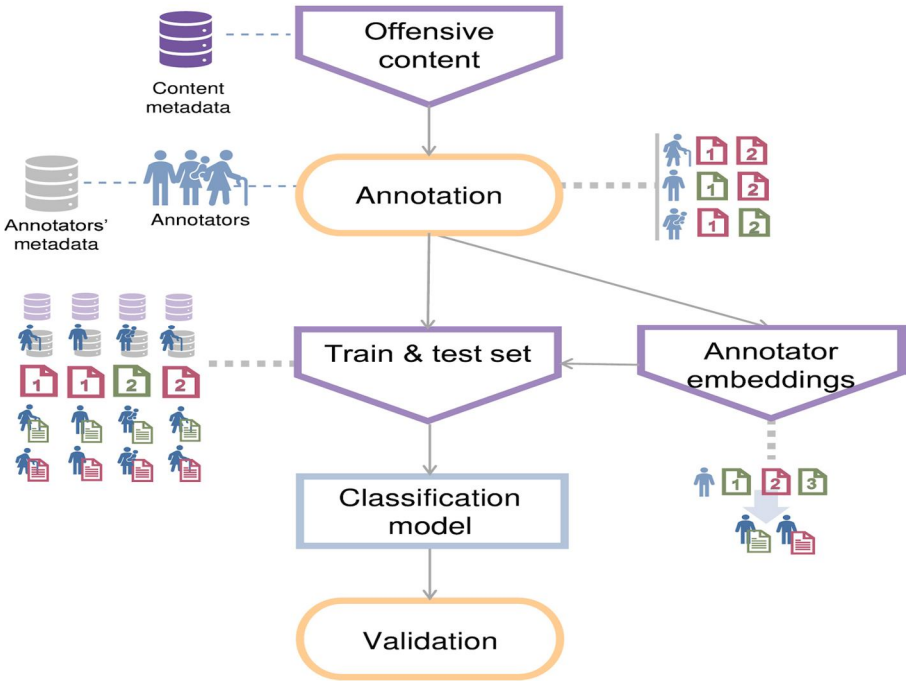


Fig16- Opinion-based Personalization

F. Performance Analysis

Here, we have compared the results of all the models we tested with.

Sentiment Analysis:

Model	Training Accuracy	Validation Accuracy	F1 score
RandomForestClassifier	99.904	95.106	59.979
Logistic Regression	98.541	94.168	59.336
DecisionTreeClassifier	99.916	93.242	53.924
SVC	97.818	95.219	49.868
XGBClassifier	94.459	94.331	35.378

We can see that DecisionTreeClassifier, gives us better Training Accuracy. But in terms of more optimal Validation Accuracy, RandomForestClassifier provides us with better accuracy.

G. Emoji Prediction

Model	Training Accuracy	Validation Accuracy
LSTM	90.799	86.682
Bi-Directional LSTM	93.74	88.74

Comparing the 2 models used, we can clearly see that Bi-Directional LSTM outperforms LSTM and gives us better results.

H. TweetEval Validation

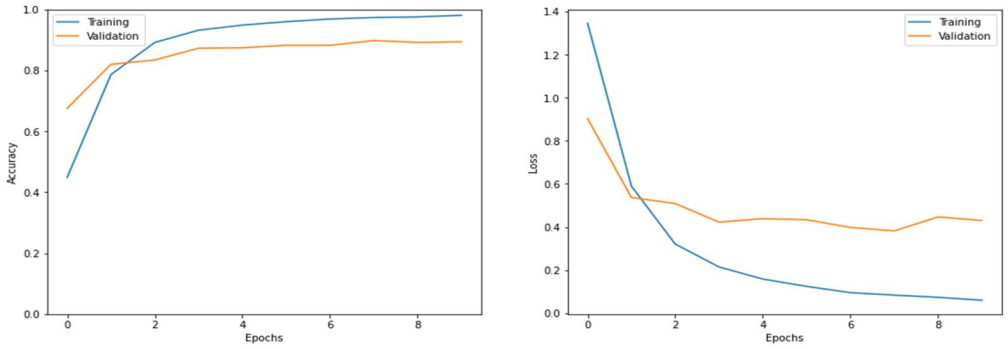
Model	Emoji	Emotion	Hate	Sentiment
SVM	24.3	63.1	71.1	68.1
FastText	23.2	62.9	71.7	59.8
BLSTM	19.3	61.5	71.8	59.6

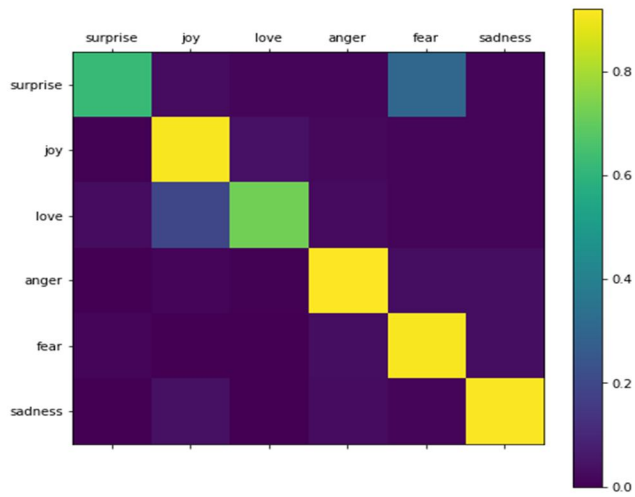
I. TweetEval Test

Model	Emoji	Emotion	Hate	Sentiment
SVM	28.5	63.8	35.7	61.1
FastText	25.2	64.9	49.3	58.3
BLSTM	23.7	64.5	51.2	56.6

The validation sets are sampled at random from the training data for those tasks where no validation split is present in the dataset.

J. Emotion Recognition using RNN





Using sequential model: Bi-directional LSTM,

Training Accuracy	Validation Accuracy
98.5	89.4

By using a separate RNN model, we are able to achieve results that are miles better than the TweetEval combined model.

IV. CONCLUSIONS

In this paper, we have introduced TwitterNLP, an NLP platform with a focus on social media. The software uses very simple language models that were trained on Twitter and adjusted for many prominent NLP tasks on social media, including sentiment analysis, identifying objectionable language, emotion recognition, emoji prediction, detecting hate speech, and named entity recognition.

Additionally, TwitterNLP makes it simple for non-programmers to analyze the models, which can assist in discovering negative biases or flaws and ultimately lead to future model improvement.

While this first released version of TwitterNLP is autarchic and complete, we want to continuously add more models and tasks to it.

We intend to create additional datasets and models for social media tasks because TwitterNLP's foundation is social media data. In particular, future expansion can be beyond the tweet categorization task, which is currently sufficed by TwitterNLP in depth.

Part-Of-Speech tagging, stopword removal, syntactic parsing has always proven challenging in a gigantic setting like social media.

In addition, we want to support languages other than English in a larger range of activities and expand TwitterNLP to include other social networking sites like Reddit, LinkedIn, and Instagram.

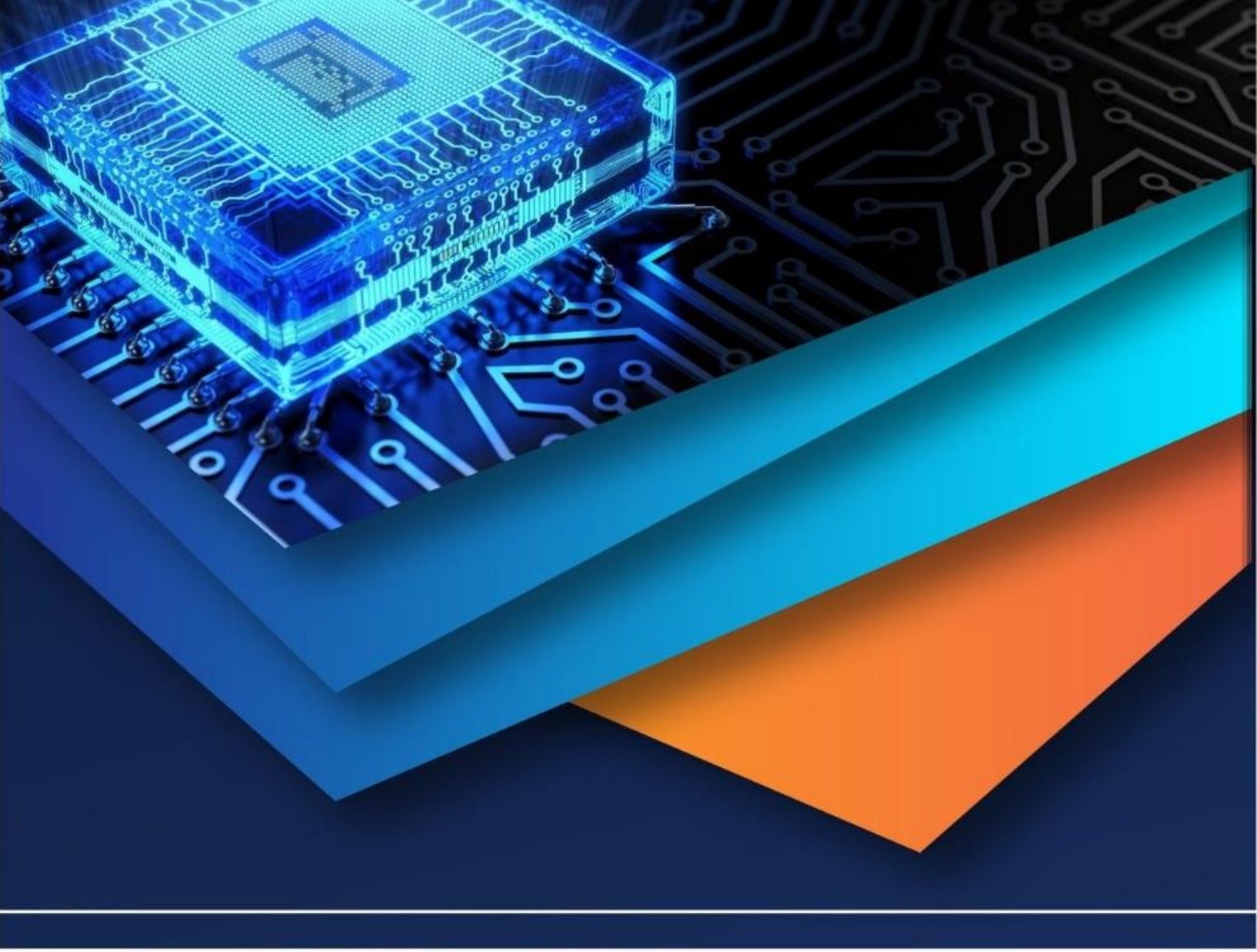
The results of this experiment indicate that TwitterNLP has multifold advantages and can be used explicitly.

REFERENCES

[1] S. M. Metev and V. P. Veiko, Laser Assisted Microtechnology, 2nd ed., R. M. Osgood, Jr., Ed. Berlin, Germany: Springer-Verlag, 1998.
[2] [TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification](https://aclanthology.org/2020.findings-emnlp.148) (Barbieri et al., Findings 2020)
[3] https://www.analyticsvidhya.com/blog/2021/06/twitter-sentiment-analysis-a-nlp-use-case-for-beginners/
[4] Pak, Alexander & Paroubek, Patrick. (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. Proceedings of LREC. 10.
[5] A. Yousaf et al., "Emotion Recognition by Textual Tweets Classification Using Voting Classifier (LR-SGD)," in IEEE Access, vol. 9, pp. 6286-6295, 2021, doi: 10.1109/ACCESS.2020.3047831.
[6] V. N. Durga Pavithra Kollipara, V. N. Hemanth Kollipara and M. D. Prakash, "Emoji Prediction from Twitter Data using Deep Learning Approach," 2021 Asian Conference on Innovation in Technology (ASIANCON), 2021, pp. 1-6, doi: 10.1109/ASIANCON51346.2021.9544680.
[7] G. A. De Souza and M. Da Costa-Abreu, "Automatic offensive language detection from Twitter data using machine learning and feature selection of metadata," 2020 International Joint Conference on Neural Networks (IJCNN), 2020, pp. 1-6, doi: 10.1109/IJCNN48605.2020.9207652.
[8] Camacho-Collados, José & Rezaee, Kiamehr & Riahi, Talayeh & Ushio, Asahi & Loureiro, Daniel & Antypas, Dimosthenis & Boisson, Joanne & Espinosa-Anke, Luis & Liu, Fangyu & Martínez-Cámara, Eugenio & Medina, Gonzalo & Buhrmann, Thomas & Neves, Leonardo & Barbieri, Francesco. (2022). TweetNLP: Cutting-Edge Natural Language Processing for Social Media. 10.48550/arXiv.2206.14774.
[9] [Feature-Rich Twitter Named Entity Recognition and Classification](https://aclanthology.org/W16-3922) (Sikdar & Gambäck, 2016).
[10] https://paperswithcode.com/dataset/hate-speech-and-offensive-language
[11] M. Krommyda, A. Rigos, K. Bouklas and A. Amditis, "Emotion detection in Twitter posts: a rule-based algorithm for annotated data acquisition," 2020 International Conference on Computational Science and Computational Intelligence (CSCI), 2020, pp. 257-262, doi: 10.1109/CSCI51800.2020.00050.
[12] https://www.ijstr.org/final-print/mar2020/Emotion-Recognition-Of-Twitter-Posts-In-Real-time-A-Survey.pdf
[13] https://www.researchgate.net/publication/339980709_Sentiment_Analysis_with_NLP_on_Twitter_Data
[14] Anupama B S , Rakshith D B , Rahul Kumar M , Navaneeth M, 2020, Real Time Twitter Sentiment Analysis using Natural Language Processing, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 09, Issue 07 (July 2020)
[15] Manju Venugopalan and Deepa Gupta, Exploring Sentiment Analysis on Twitter Data, IEEE 2015
[16] Kishori K. Pawar, Pukhraj P Shrishrimal, R. R. Deshmukh, Twitter Sentiment Analysis: A Review International Journal of Scientific & Engineering Research, Volume 6, Issue 4, April-2015
[17] B. Pariyani, K. Shah, M. Shah, T. Vyas and S. Degadwala, "Hate Speech Detection in Twitter using Natural Language Processing," 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), 2021, pp. 1146-1152, doi: 10.1109/ICICV50876.2021.9388496.
[18] Pitsilis, Georgios & Ramampiaro, Heri & Langseth, Helge. (2018). Effective hate-speech detection in Twitter data using recurrent neural networks. Applied Intelligence. 48. in press.. 10.1007/s10489-018-1242-y.
[19] Park, Ji & Fung, Pascale. (2017). One-step and Two-step Classification for Abusive Language Detection on Twitter.

[20] V. N. Durga Pavithra Kollipara, V. N. Hemanth Kollipara and M. D. Prakash, "Emoji Prediction from Twitter Data using Deep Learning Approach," 2021 Asian Conference on Innovation in Technology (ASIANCON), 2021, pp. 1-6, doi: 10.1109/ASIANCON51346.2021.9544680.
[21] Wolny, Wieslaw. (2016). TWITTER SENTIMENT ANALYSIS USING EMOTICONS AND EMOJI IDEOGRAMS.
[22] Pitsilis, Georgios & Ramampiaro, Heri & Langseth, Helge. (2018). Detecting Offensive Language in Tweets Using Deep Learning.
[23] M. Kanakaraj and R. M. R. Guddeti, "NLP based sentiment analysis on Twitter data using ensemble classifiers," 2015 3rd International Conference on Signal Processing, Communication and Networking (ICSCN), 2015, pp. 1-5, doi: 10.1109/ICSCN.2015.7219856.
[24] Garg, Y., Chatterjee, N. (2014). Sentiment Analysis of Twitter Feeds. In: Srinivasa, S., Mehta, S. (eds) Big Data Analytics. BDA 2014. Lecture Notes in Computer Science, vol 8883. Springer, Cham.
[25] TWEETEVAL: Unified Benchmark and Comparative Evaluation for Tweet Classification Francesco Barbieri, Jose Camacho-Collados Leonardo Neves, Luis Espinosa-Anke Snap Inc., Santa Monica, CA 90405, USA School of Computer Science and Informatics, Cardiff University, United Kingdom.
[26] [TimeLMs: Diachronic Language Models from Twitter](https://aclanthology.org/2022.acl-demo.25) (Loureiro et al., ACL 2022)
[27] [T-NER: An All-Round Python Library for Transformer-based Named Entity Recognition](https://aclanthology.org/2021.eacl-demos.7) (Ushio & Camacho-Collados, EACL 2021)
[28] https://paperswithcode.com/paper/xlm-t-a-multilingual-language-model-toolkit
[29] Kalyan KS, Rajasekharan A, Sangeetha S. AMMU: A survey of transformer-based biomedical pretrained language models. J Biomed Inform. 2022 Feb;126:103982. doi: 10.1016/j.jbi.2021.103982. Epub 2021 Dec 31. PMID: 34974190.
[30] Yang F, Wang X, Ma H, Li J. Transformers-sklearn: a toolkit for medical language understanding with transformer-based models. BMC Med Inform Decis Mak. 2021 Jul 30;21(Suppl 2):90. doi: 10.1186/s12911-021-01459-0. PMID: 34330244; PMCID: PMC8323195.
[31] https://www.kaggle.com/code/mangipudiprashanth/twitter-sentiment-analysis-using-ml-nlp

©IJRASET: All Rights are Reserved



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)