The use of hate speech is widespread all around us, due to unprejudiced opinions forming overtime, in accordance with religion, culture or gender. In the last decade or two the use of hate speech has been normalized to some degree, with the introduction of online social media platforms. These platforms allow its users freedom of speech which can produce a negative knockback on the users. Due to this reason the paper by Haroon Shakeel on hate speech and offensive language detection in roman urdu, aimed to further the research in this area. The paper hypothesized that previously published research may have focused on detection of racism and sexism, and other sectors differentiating offensive language and hate speech, diversifying the area of study which may have given rise to problems, such as duplication and reduction of interrelationships, encouraging decreased usability across different sectors of this study. These issues have been targeted, as models for hate speech detection in a diverse pool of languages have formed under the name of Hate Speech and Offensive Content Identification (HASOC). In recent years multiple datasets have been made available to the public in several languages. Even so there is no publicly available dataset for Roman Udru to the best of our knowledge.

Roman Urdu lacks language resources and datasets. Catering to these issues the research presents a lexicon of hateful words coving multivariable groups catering to abusive and derogatory terminology with slurs and terms connotating to sexist language and religious hate, with separately defined Roman Urdu commonly used word sets in the language developing annotated datasets with labels of offensive language utilizing tweets to create these models. One might consider this a contradiction and an inefficient base for the paper as manual reviews are performed by most social media platforms in order to address these issues. But in a long term perspective these methods can become obsolete and inefficient, as the reviews depend on reviewer speed and current literature knowledge and understanding.

Following the updating and reforming of the lexicons sampling for annotations. The research moves towards the creation of a golden standard, under the name of Roman Urdu Hate Speech and Offensive Language Detection ( RUHSOLD ) datasets. The gold-standard was developed for coarse grained classifications as the first subtask referring to hate offensive and normal content. On the other hand the demographic of users who may converse in roman urdu are separated in a different subtask called fine grain classification. The division of the golden standard allows the researchers to approach scenarios with differing difficulties separately. The labels for the golden standard include, Abusive/offensive, sexism, religious hate, profane and normal. The data itself is divided into train,test and validation using the stratified sampling method. Completing the RUHSOLD dataset which was made public to further the research perspectives.

Once the dataset is made using the RUHSOLD the experimental design evaluates the performance of different embeddings, baseline models and proposed models for both classifications. Using the black box method the out of the box performance of multilingual embeddings for which the data was collected using twitter API are compared and fine tuned on RUHSOLD gauging the transfer learning capability onto a different domain and language. COming towards the baseline models were selected based on their reported performance for hate speech detection on multiple datasets for reimplementation form the companion codes.

The paper focuses on the use of end to end deep learning, training and upgrading current state of the art automatic speech recognition (ASR) pipelines, which is a powerful alternative solution replacing most modules with a single model allowing speaker adaptation and complex feather extraction, in addition to allowing train and capture large sets and model high performance computing. The paper further explores multiple network architectures diving into bidirectional and unidirectional models, evaluating RNNs with greater strides and bigram outputs for English, using SortaGrad and Batch Normalization for numerical optimization. This investigation was made possible by a highly efficient, HPC-inspired training method that enables quick training of new, full-scale models on a massive datasets with reduced error rates and increasing accuracy.

Using Deep speech 2 (DS2) for speech recognition, researchers have focused on three key factors: model architecture, sizable labeled training datasets, and computational scale. They have also thoroughly examined model architectures and the impact of data and model size on recognition performance.An internal datasets for English (3,600 hours) and Mandarin (1,400 hours) were built using raw data that was originally recorded as audio segments with slurred transcriptions. These films vary in length from several minutes to more than an hour, making it impossible to unroll them in the RNN in a timely manner while it is being trained. To address this, we created an alignment, segmentation, and filtering process that can provide a training set with fewer incorrect transcriptions and shorter utterances. Research results revealed that Deep Speech, when compared to its previous iteration, has significantly narrowed the performance gap in transcription with human workers by utilizing more data and larger models in addition to simplifying application to new languages, indicating that the goal of a single speech system that outperforms humans in most scenarios is quickly approaching.

**Deep Speech: Scaling up end-to-end speech recognition**

The main objective was to enhance performance in noisy situations, where current systems fail.

The researchers assembled a dataset consisting of 5000 hours of speech from 9600 speakers, categorized according to datasets and type, read or conversational. The first alternative, capturing labeled data, is not practicable, so the researchers devised a method involving the creation of a larger number of shorter clips, which are then processed individually and treated as unique sources of noise. Then, the audio signals are created by superimposing source signals to create a synthetic version of noisy training data. Where necessary, reverberation, echoes, and other types of damping are added to the audio signals, which are then combined to create an audio picture that is pretty realistic. The "Lombard Effect'' was a difficult effect for voice recognition systems to deal with so the effect is purposefully induced during data collection by playing loud background noise through a person's headphones when they record an utterance. This effect is then represented in the training data by the loud background noise. The noise causes them to inflict their speech, which enables the Lombard effect to be captured in the training data. Hub5'00 was used to assess the system's performance against earlier research using a recognised but extremely difficult test set. The Deep Speech system improves on this baseline by 2.4% absolute WER and 13.0% relative when trained on the total 2300 hours of data. The Maas et al. model (DNN-HMM FSH) achieves 19.9% WER on the Fisher 2000 hour corpus. That system was created using the cutting-edge open source speech recognition programme Kaldi. This finding was included in the study to show that Deep Speech can compete with the top ASR systems. Two RNNs were trained, one on 5000 hours of raw data and the other on the same 5000 hours plus noise. The noisy model outperforms the clean model by 22.6% WER on the 100 noisy utterances. This is a 6.1% absolute and 21.3% relative improvement. Combining these approaches allowed the researchers to create a data-driven speech system that is both more effective than current approaches and does not rely on the intricate processing stages that had stalled future development.

**Deep Speech 2: End-to-End Speech Recognition in English and Mandarin**

On a test set of 1882 samples of loud speech and a development set of 2000 utterances. The results demonstrate that the deepest model with 2D-invariant 19 convolution and BatchNorm outperforms the shallow RNN by 48% relative, continuing the trend started with the English system—multiple layers of bidirectional recurrence significantly improve performance. So the researchers were able to deploy the system at low latency and high throughput without sacrificing much accuracy, on a held-out set of 2000 utterances. The system achieved 5.81 character error rate whereas the deployed system achieved 6.10 character error rate. This is only a 5% relative degradation for 22 the deployed system. In order to accomplish this, they employed a neural network architecture with low deployment latency, reducing the precision of the network to 16-bit, building a batching scheduler to more efficiently evaluate RNNs, and finding a simple heuristic to reduce the beam search cost. These techniques allowed the deployment of Deep Speech at low costs for interactive applications.