1. Punctuation restoration is the task of adding punctuation symbols to raw text.
2. Light stemming refers to a process of stripping off a small set of prefixes and/or suffixes, without trying to deal with infixes, or recognize patterns and find roots.
3. Recursive suffix stripping is a recursive process to eliminate multiple suffixes from a single word.
4. In feed forward neural network, the flow of information takes place in the forward direction, as x is used to calculate some intermediate function in the hidden layer which in turn is used to calculate y.
5. Margin infused relaxed algorithm is designed to learn a set of parameters by processing all the given training examples one-by-one and updating the parameters according to each training example, so that the current training example is classified correctly with a margin against incorrect classifications at least as large as their loss.
6. The principle of maximum entropy states that the probability distribution which best represents the current state of knowledge about a system is the one with largest entropy.
7. The Bangla-English MT (machine translation) system works by analyzing prepositions for both Bangla and English.
8. Example based MT system translates from Source Language to Target Language in the chunk and generate with morphological analysis with the help of WordNet.
9. Active learning can improve the quality of a phrase-based SMT system, and show significant improvements in translation compared to a random sentence selection baseline, when test and training data are taken from the same or different domains
10. LSTM, CNN, RNN
11. Different n-grams results in terms of frequency, normalized frequency, and ranked frequency. It is seen that when n is increased from 1 to 3, the performance increases. However, from a value of 3 to 4 or more, the performance decreases.
12. BERT, DistilBERT, RoBERTa, and XLM-RoBERTa.
13. Some fraction of the input tokens are randomly masked, and the objective is to predict the vocabulary ID of the original token in that position. The bidirectional nature ensures that the model can effectively use both past and future contexts for this task.
14. This is a binary classification task where given two sentences, the goal is to decide whether the second sentence immediately follows the first sentence in the original text. Positive sentences are created by taking consecutive sentences from the text, and negative sentences are created by taking sentences from two different documents.
15. The multilingual variant of BERT (mBERT) is trained using the Wikipedia corpus of the most extensive languages. Data is sampled using an exponentially smoothed weighting to address differences among the corpus size of different languages, ensuring that high resource languages like English are under-sampled compared to low resource languages. Word counts are weighted similarly so that words from low-resource languages are represented adequately in terms of vocabulary.
16. RoBERTa improves upon BERT by proposing several novel training strategies, including (1) training the model longer with more data (2) using a larger batch size (3) removing the next sentence prediction task and only using MLM loss (4) training on longer sequences (5) generating the masking pattern dynamically. These modifications allow RoBERTa to outperform BERT on different downstream language understanding tasks consistently.
17. Electra is trained using a sample-efficient pre-training task called replaced token detection. In this approach, instead of masking input tokens, they are replaced with alternatives sampled from a generator network. Then a discriminator model is trained to predict whether a generator sample replaced each token or not. This approach allows the model to learn better representation while being compute-efficient
18. Sentiment, emotion, authorship attribution, and news categorization.
19. BNLP Toolkit
20. It is used to segment the text into subword units and applied byte pair encoding to increase the consistency of data segmentation for handling rare words
21. It is used to binarize the data for training.
22. Adam optimizer involves a combination of two gradient descent methodologies: Momentum & Root Mean Square Propagation (RMSP)
23. It shows higher performance because it is a large version of the model, whereas monolingual models are the base version, consisting of fewer parameters than the larger version.
24. Different types of inputs i.e. token only and with additional inputs, produce different outputs thus affecting the performance of the models used.
25. Monolingual models did not perform well for this task. The results showed that the XLM-RoBERTa large model performs the best across different datasets such as news, manual, and ASR transcriptions. Among the monolingual models, Indic-DistilBERT performs better overall. For news text, manual, and ASR transcriptions, the F1 score is 80%, 61.5%, and 58.0%, respectively. As expected, the performance on the news test set is better than the transcribed texts for all models. Due to errors introduced by the ASR model, performance in ASR transcriptions are lower than manual transcriptions. Among different labels, performance in Comma is significantly worse in the transcribed texts.
26. Transformer-based model takes significantly high computation time even during inference
27. Comparatively better results can be obtained using the XLM-RoBERTa model compared to others. Across multilingual transformer models, DistilBERT-m is the second best model. Among the monolingual models, the Indic-RoBERTa model performs better than other models for emotion classification.
28. FastText embedding trained on different corpora was used to train the KNN classifier.
29. Indic-BERT, Indic-XLM-RoBERTa, XLM-RoBERTa, BERT-bn, BERT-m were better suited for news categorization task.
30. (i) the availability and accessibility of resources; (ii) the lack of reliability on annotated labels, with no proper guidelines or evaluation measures like inter-annotator agreement; and (iii) the lack of comparable benchmarking performance – due to the absence of well-defined train/dev/test splits
31. It is trained on a larger corpus and has a larger vocabulary size. As a result, fewer unknown tokens are introduced after tokenization.
32. As XLM-RoBERTa is a model trained with more layers and parameters, it also has better generalization ability compared to other models. This results in better performance in the downstream tasks.