Checkpoint 1 | Project Outline

Scope

This project involves a study regarding the classification of text using character-level models and how they compare with other baseline models. Classification tasks have always been one of the most important tasks. A classification task is a type of machine learning or data analysis problem where the goal is to categorize or assign predefined labels or classes to input data based on certain features or characteristics. The primary objective of a classification task is to build a model that can learn from a labelled dataset and then use this learned knowledge to classify new, unseen data into one of the predefined classes or categories.

There are three different datasets the classification needs to be performed on: Yelp Review Full, dbpedia 14, Amazon polarity, which are of the form: content, title, and class. The models to be used are:

- Baseline
  - BoW
  - BoW with TF-IDF
  - LSTM
- Character Level Models
  - Character Level Convolution Network
  - CANINE

With the help of the baseline models, initial classification can be done on the datasets and the points where errors occur can be discovered and later the better models can be implemented to try and reduce the error cases.

Our aim is to try to build the baseline models, and make them perform to the best levels possible, highlight the places of failure, then make the improved models to tackle the initial failure points, while showing how and where it tackles and fails as well and at the end, develop an end-to-end system.

Literature Survey

Baseline Models

BoW The "Bag of Words" (BoW) is a fundamental technique used in Natural Language Processing (NLP) for text analysis and text classification tasks. It represents text data as a collection of individual words, ignoring their order and structure but keeping track of their frequency in the document. The basic idea is to create a "bag" containing all the unique words (or tokens) in a corpus of text, along with their respective counts or frequencies.

Formula:
The BoW vector of document d is a vector of length V where the value at each index i represents the frequency of word i in document d.

$$\text{BoW}(d) = [\text{wordcount}(d, 1), \text{wordcount}(d, 2), \ldots, \text{wordcount}(d, V)]$$

Where V is the size of the vocabulary, d be a specific document, wordcount(d, w) be the count of word w in document d.

BoW with TF-IDF The Bag of Words (BoW) model, combined with TF-IDF (Term Frequency-Inverse Document Frequency), is a common approach in Natural Language Processing (NLP) to represent and analyse text data. It enhances the basic BoW model by considering the importance of words in a document relative to their frequency in the entire corpus of documents. TF-IDF is used to assign weights to words in the BoW representation, allowing it to capture the significance of words in each document.

Formula:
$TF(w, d)$ = number of times word w occurs in document d / total number of words in document d

$$IDF(w) = \log\left(\frac{\text{total number of documents}}{\text{number of documents with word } w + 1}\right)$$

$TF - IDF(w, d) = TF(w, d) * IDF(w)$

We can use the above formula of TF-IDF in BoW.

LSTM Long Short-Term Memory (LSTM) is a type of recurrent neural network (RNN) architecture that is commonly used in Natural Language Processing (NLP) and other sequence modelling tasks. LSTM is designed to address the vanishing gradient problem that traditional RNNs face, allowing it to capture long-range dependencies in sequential data, which is essential for understanding and generating natural language text.

Using forget gate, input gate and output gate, hidden states, cell states and outputs are generated which are used in LSTMs.

Character-Level Representations

Character-level representations have been explored in various NLP tasks. Models like character-level Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks have shown promise in capturing character-level features.

Convolutional Neural Networks (CNNs) for Text Classification From the paper Character-level Convolutional Networks for Text Classification

CNNs have demonstrated effectiveness in text classification tasks. The Kim CNN model, for instance, employs convolutional layers to extract meaningful features from text at different granularities. However, these models typically operate at the word level.

Zhang et al.'s Character-Level CNN Model The paper introduces a novel approach to text classification by processing text at the character level. Their model employs convolutional layers to extract character-level features and has shown promising results, especially for languages with rich morphology.

CANINE: A Character Level Transformer From the paper CANINE: Pre-training an Efficient Tokenization-Free Encoder for Language Representation.

CANINE (Character Architecture with No tokenization In Neural Encoders) is a neural encoder that operates directly on character sequences, without explicit tokenization or vocabulary, and a pre-training strategy that operates either directly on characters or optionally uses subwords as a soft inductive bias. To use its finer-grained input effectively and efficiently, CANINE combines downsampling, which reduces the input sequence length, with a deep transformer stack, which encodes context.

Advantages of Character-Level Models Character-level models offer several advantages, including:

- Handling out-of-vocabulary words effectively.
- Capturing morphological information.
- Reducing data pre-processing requirements.

Comparison with Word-Level Models Character-level models need to be compared to word-level models in terms of:

- Performance in various text classification tasks.
- Computational efficiency.
- Robustness to noisy or misspelled words.

Applications and Use Cases Character-level models find applications in:

- Sentiment analysis.
- Authorship attribution.
- Language identification.
- Multilingual text classification.

Conclusion Character-level convolutional networks for text classification represent a promising avenue in NLP, offering solutions to some of the challenges posed by word-level models. Understanding their advantages, limitations, and potential applications is crucial for advancing text classification techniques.

Project Timeline

- By Oct 7: Outline (**Checkpoint 1**)
- By Oct 15: Completion of model BoW
- By Oct 22: Completion of model BoW with TF-IDF

- By Oct 28: Completion of LSTM model
- By Oct 30: Completion of Baseline models (**Checkpoint 2**)
- By Nov 13: Completion of model on Character Level Convolution Network
- By Nov 18: Finetuning on CANINE
- By Nov 25: Intuitive and Explanatory Analysis, Results and Limitations for all the models
- By Nov 27: Completion of end-to-end system (**Checkpoint 3**)