



Email Author Identification

SMAI Project, 2017

*Dr. Vineet Gandhi
Ishit Mehta*

TEAM 15

*Aditya Srivastava
Karthik Chintapalli
Kritika Prakash
Vighnesh Chenthil Kumar*



Objective

- Classify e-mails from the **ENRON dataset** based on their predicted authors
 - ~500,000 emails
 - 153 unique authors - employees of ENRON
 - Released into the public domain after an investigation led to the closure of the company
- Identify authors of test emails based on model created



Approach

- Familiarizing with existing state-of-the-art document classification techniques
- Studying existing approaches to ENRON specific document classification
- Formalizing the final approach and further experiments to try and improve upon the existing models



Phase 1: Data Preparation

- Fixed the number of author to allow more experimentation with features
- Number chosen such that
 - the number of emails per author is maximized
 - emails/author ratio is similar across all authors
- This value was found to be **10** authors with **800-1000** emails each



Phase 1: Data Preparation

- Cleaning the corpus
 - The ENRON data consists of raw, unparsed, unclean e-mails
 - Extracting only the **Subject** and **Body** for every email
 - Removing chains of forwarded messages
 - **Word, sentence** and **paragraph**-level tokenization
 - Case normalization



Phase 2: Existing Neural Models

- Explore the state-of the art techniques in document classification and implement them on our dataset
- The models implemented include
 - **CNN**-based model
 - **Bi-LSTM** based model
 - **Hierarchical Bi-LSTM** model

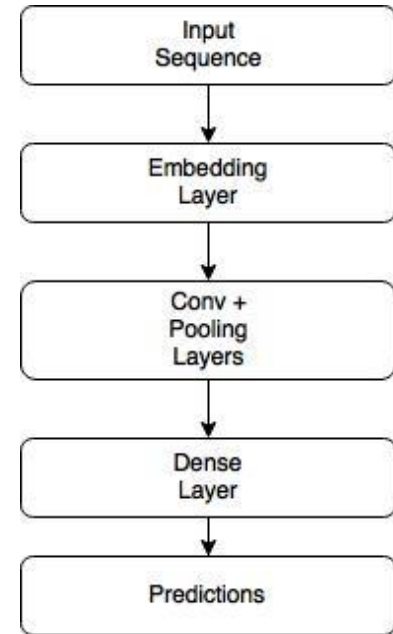


Phase 2.1: CNN-based Model

- Implementation of ideas from “*Convolutional Neural Networks for Sentence Classification*” - Yoon Kim
- CNNs are known to capture localized chunks of information - this can be useful to find phrasal units within long texts
- Can identify key, commonly used groups of words by an author

Phase 2.1: CNN-based Model

- The **Embedding layer** generates a sequence of word-embeddings from a sequence of words.
- Each **Conv layer** has 128 5x5 filters
- The **Dense layer** is used for classification



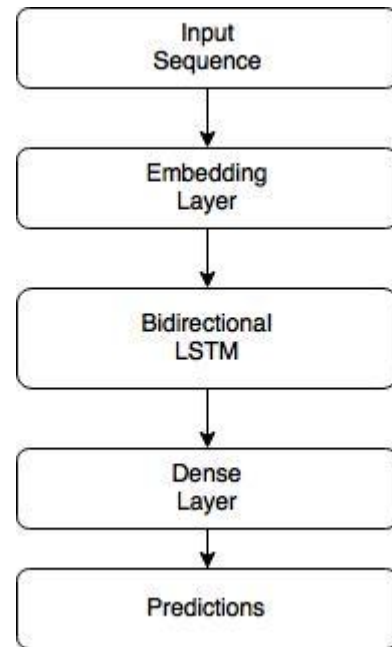


Phase 2.2: **Bi-LSTM-based Model**

- Standard, commonly used technique in text classification
- **LSTM**s are a special kind of **RNN** which are more capable of remembering long term dependencies in a sequence
- Helps in author classification as it keeps track of past events while processing a sequence of text

Phase 2.2: Bi-LSTM-based Model

- The **Embedding layer** generates a sequence of word-embeddings from a sequence of words
- The **Bidirectional LSTM** generates e-mail embeddings from the sequence of word embeddings
- The **Dense layer** is for the classification



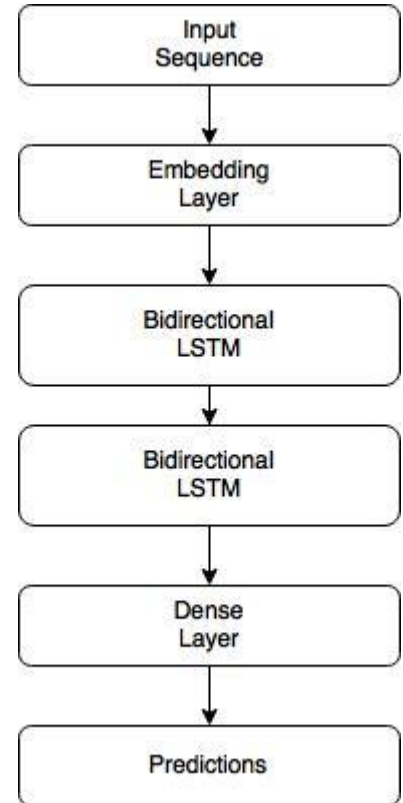


Phase 2.3: Hierarchical Bi-LSTM Model

- Idea based on “*Hierarchical Attention Networks for Document Classification*” - Yang *et. al.*
- LSTMs are known to work best for a sequence length of 10-15 elements
- In a Hierarchical LSTM,
 - The first level generates sentence-embeddings from word-embeddings
 - The second level generates document-embeddings from sentence-embeddings

Phase 2.3: Hierarchical Bi-LSTM Model

- The **Embedding layer** generates a sequence of word-embeddings from a sequence of words
- The **first Bidirectional LSTM** generates sentence embeddings from the sequence of word embeddings
- The **second Bidirectional LSTM** generates e-mail embeddings from sentence embeddings
- The **Dense layer** is for the classification





Phase 2.4: Results

- The best validation accuracies obtained for each of the models are given below

CNN model	56.43%
Bi-LSTM	73.54%
Hierarchical Bi-LSTM	77.04 %

- As anticipated, the Hierarchical Bi-LSTM model performs the best



Phase 3: **Stylometric Features**

- Models tried so far are generally useful for document classification tasks
- Emails stand apart from documents in two major ways
 - Short text length
 - Lack of topical difference
- So, existing work specific to the ENRON dataset rely on stylometric features.
- Stylometric features capture certain writing styles specific to authors
- We attempt to augment our Hierarchical Bi-LSTM model with these stylometric in an attempt to improve performance.

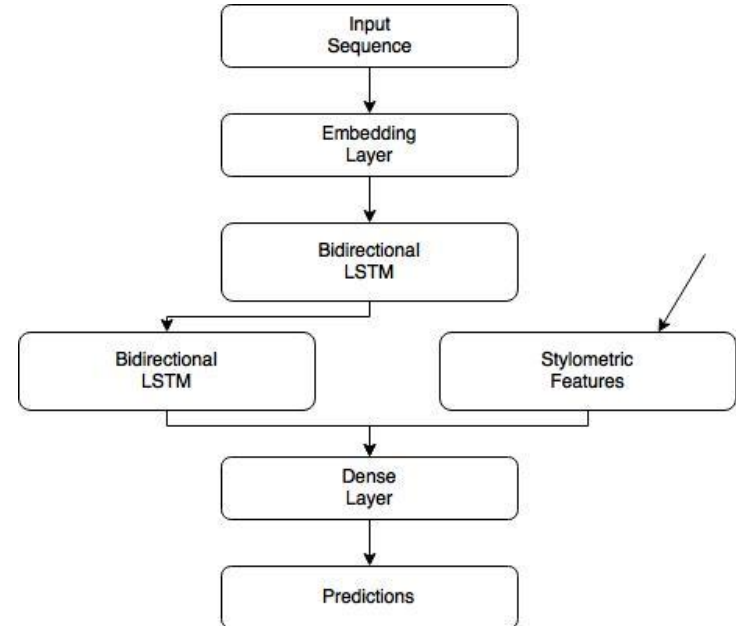


Phase 3.1: Stylometric Feature Extraction

- The following stylometric features were extracted per author per e-mail:
- **Lexical:**
 - Average sentence-length
 - Average word-length
 - Total number of words
 - Ratio of unique words to total number of words
 - Total number of characters
- **Syntactic:**
 - Total number of function words
 - Total number of personal pronouns
 - Total number of adjectives

Phase 3.2: Augmenting Hierarchical Bi-LSTM Model

- The **stylometric features** are appended in the Hierarchical Bi-LSTM model, before the final document embedding is passed on to the Dense layer for classification
- The classification is then done on these augmented document-embeddings





Phase 3.3: Final Results

- The most discriminating features were found to be
 - Ratio of unique words to total words
 - Number of adjectives
 - Average word-length
 - Average sentence-length
- The best validation accuracy obtained was ~78% (~1% increase over the basic Hierarchical Bi-LSTM model)



References:

- [CEAI: CCM-based email authorship identification model](#) - Serwat Nizamani, Nasrullah
- [Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace](#) - Ahmed Abbasi, Hsinchun Chen
- [Detection of Fraudulent Emails by Authorship Extraction](#) - A. Pandian, Mohamed Abdul Karim



Thank you!