Classification of Legal Text Using Deep Learning: Evaluation of General-Purpose Resources for a Legal Domain-Specific Task

Master Thesis Defense

Jay Vala

Otto-von-Guericke-University Magdeburg
Faculty of Computer Sciences

June 25, 2019

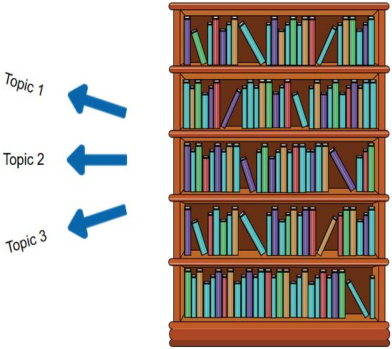
Agenda

Motivation

- Large availability of digital legal corpora
  - Legislative Text
  - Case Laws
  - International Treaties
- These laws come from different legislations, have different style, structure and language[1].
- No coherent way of organization and retrieval.
- Different vocabulary which changes with time, context and authority[1].

Motivation

- Topic-based classifier
- State-of-art solution - Support Vector Machines (SVM)



- SVMs work well for semi-structured and unstructured data.
- Have kernels for non-linear relations.
- But,
  - Parameter Sensitivity[2].
  - TF-IDF(loss of semantic/syntactics)[3].
  - Scalability issue.

Motivation

Topic-based classifier - Artificial Neural Network[4]

- Noise - No single representation of the law.
- Poorly understood intrinsic structure - No single officialdom knows every law.
- Changing characteristics - Law changes frequently.
- Scalable
- Word Embedding - considers syntactics.

Goals

- Investigate popular machine learning (SVM) and deep learning (BiLSTM) algorithms with different configurations on EUR-Lex summaries.
- Examine the performance of general-purpose resources for a legal domain-specific task.
- Look into performance benefits of multilingual data that is widely available in legal domain.
- Exploring viability of these algorithms in domain-specific settings.

Background

Document Categorization

- Analysis and assignment of documents in some predefined categories.
- Helps in retrieval of documents based on search query.
- Manual assignment is not feasible due to continuous addition of new documents every day.

Background

Natural Language Processing (NLP)

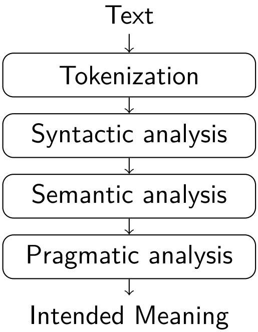


Figure: Stages of processing natural language[5]

- Sub-field of Artificial Intelligence.
- Natural language is converted into numbers and computer finds patterns.
- These patterns can be used for text summarizing, categorization.
- Stages of NLP,
  - Tokenization - dividing text into words or characters.
  - Syntactic analysis - order and structure of text
  - Semantic and pragmatic analysis - meaning and context respectively.

Background

- LSTM - Long Short Term Memory
- Variant of Recurrent Neural Network
- Have recurrent connection.
- Effective for sequence data - text, time-series, videos.
- Bidirectional LSTM - process sequence in both direction (forward and backward)

Approach and Implementation

- **Research Question 1**

Can Bidirectional Long Short Term Memory (BiLSTM) achieve better results in terms of the evaluation metrics (Precision, Recall, and F-Score) than the thoroughly studied text classification methods such as Support Vector Machines?

Approach and Implementation

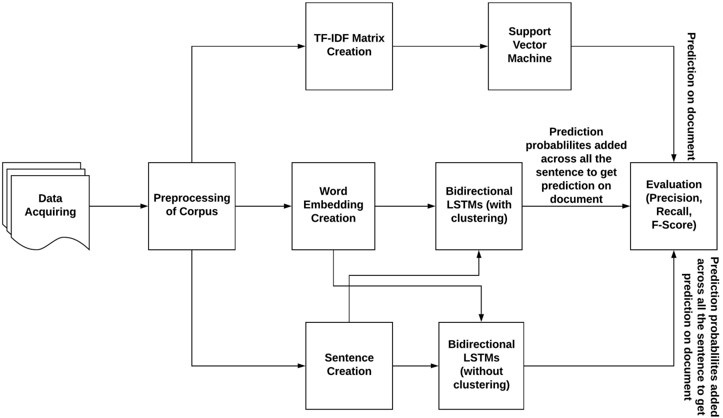


Figure: Flow chart representing the work-flow for the first research question

Approach and Implementation

- **Data Acquisition and preprocessing**

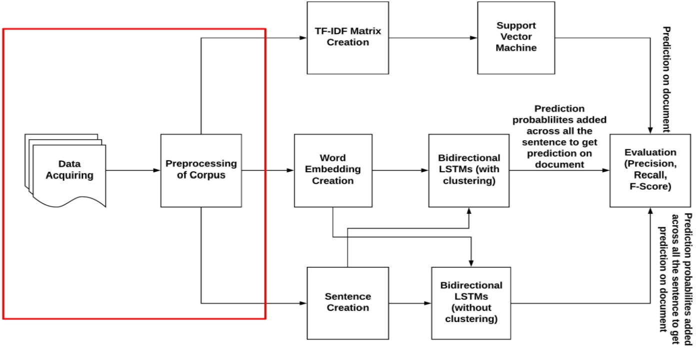


Figure: Flow chart representing the work-flow for the first research question

Approach and Implementation

- **Feature Engineering**

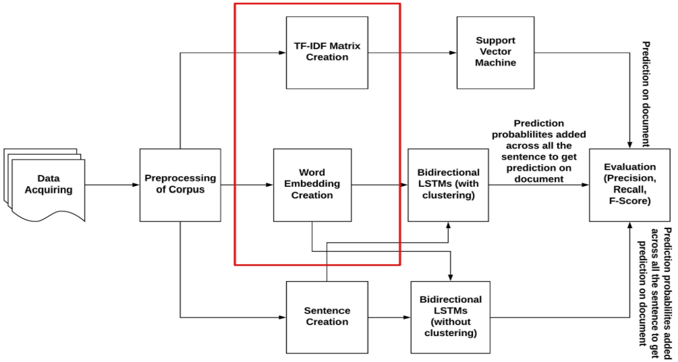


Figure: Flow chart representing the work-flow for the first research question

Approach and Implementation

- **Sentence Creation**

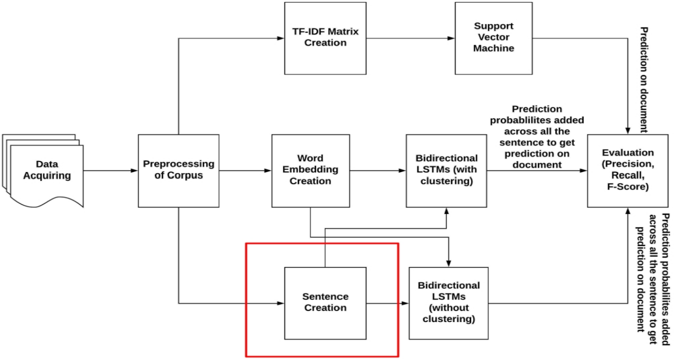


Figure: Flow chart representing the work-flow for the first research question

Approach and Implementation

- **Classifier Training**

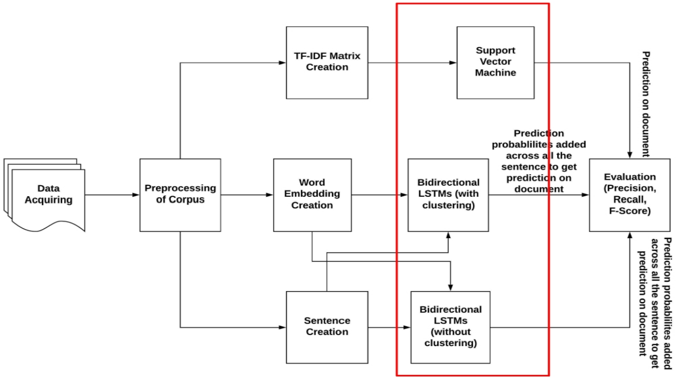


Figure: Flow chart representing the work-flow for the first research question

Approach and Implementation

- **Evaluation**

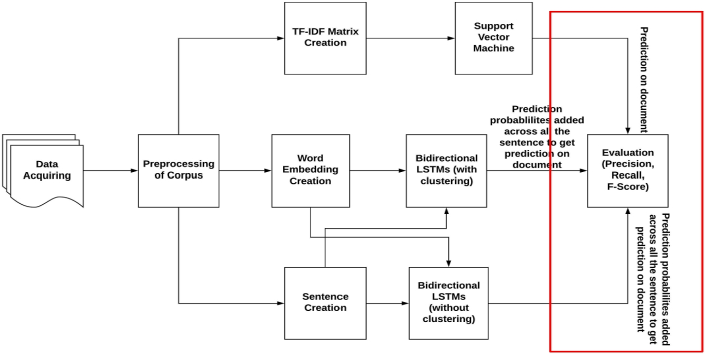


Figure: Flow chart representing the work-flow for the first research question

Approach and Implementation

- **Research Question 2**

Are general-purpose resources such as pre-trained word embeddings in case of BiLSTM applicable to specific legal domain tasks in terms of evaluation metrics? Also, further training them on legal corpus, produces comparable results to the ones only trained on legal texts?

Approach and Implementation

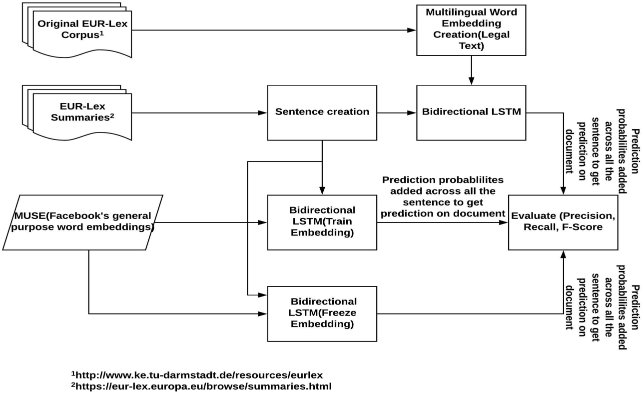


Figure: Flow chart representing the work-flow for the second research question

Approach and Implementation

- **Data Acquisition**

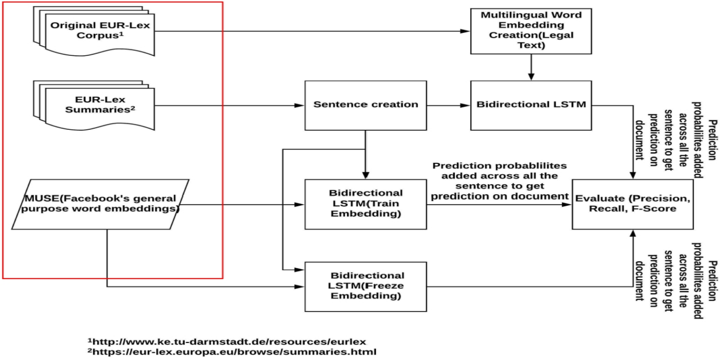


Figure: Flow chart representing the work-flow for the second research question

Approach and Implementation

- **Feature Engineering**

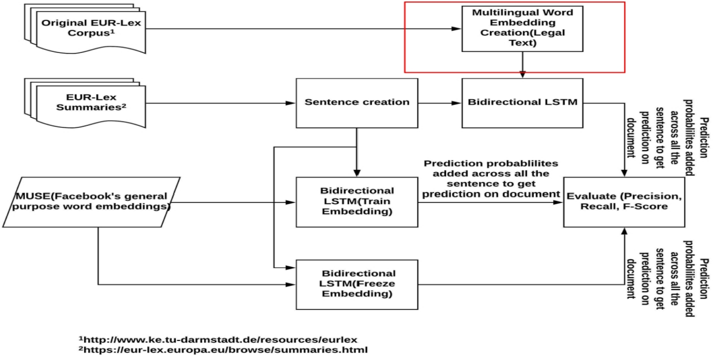


Figure: Flow chart representing the work-flow for the second research question

Approach and Implementation

- **Sentence Creation**

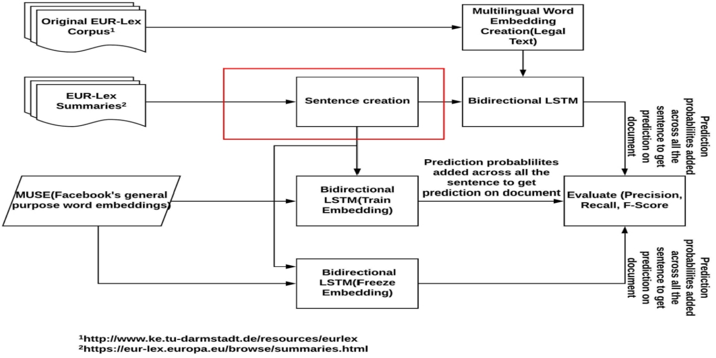


Figure: Flow chart representing the work-flow for the second research question

Approach and Implementation

- **Classifier Training**

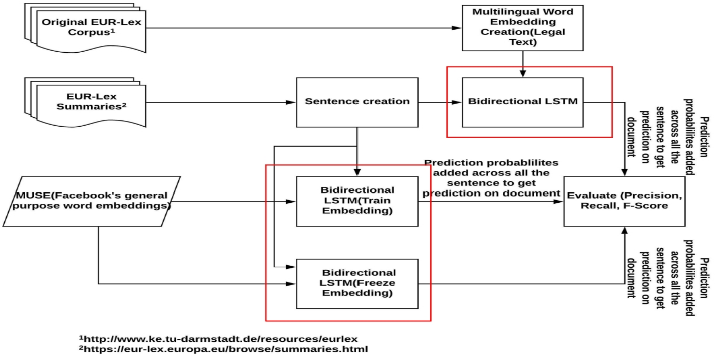


Figure: Flow chart representing the work-flow for the second research question

Approach and Implementation

- **Evaluation**

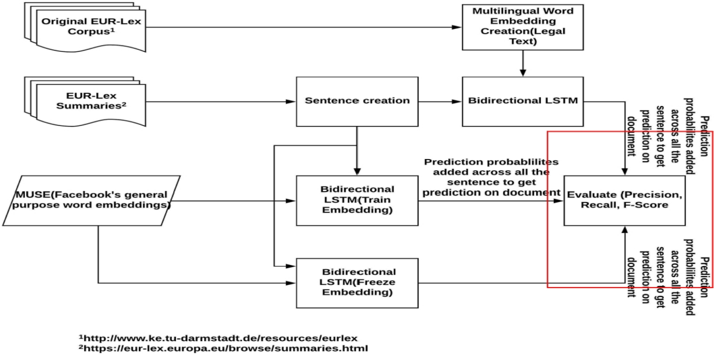


Figure: Flow chart representing the work-flow for the second research question

Approach and Implementation

- **Research Question 3**

Can BiLSTM perform better when training multiple languages in a single model, compared to training one model for each language separately?

Approach and Implementation

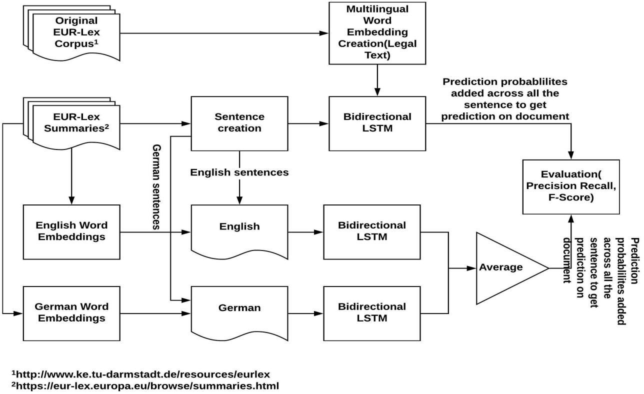


Figure: Flow chart representing the work-flow for the third research question

Approach and Implementation

- **Data Acquisition**

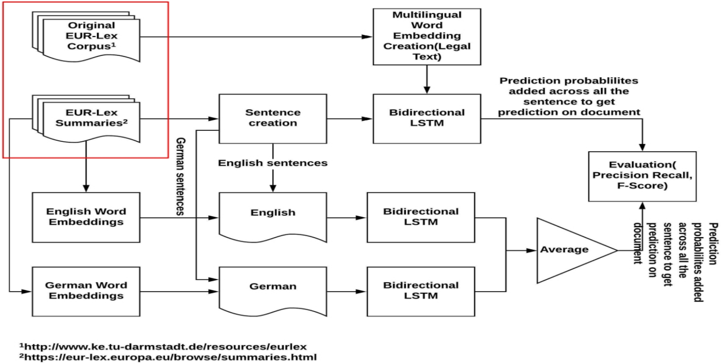


Figure: Flow chart representing the work-flow for the third research question

Approach and Implementation

- **Feature Engineering**

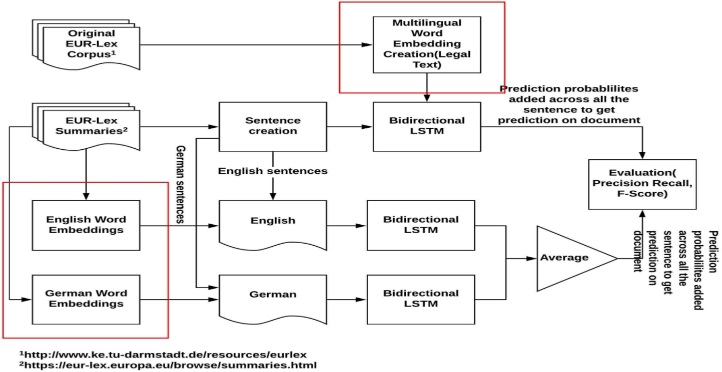


Figure: Flow chart representing the work-flow for the third research question

Approach and Implementation

- **Sentence Creation**

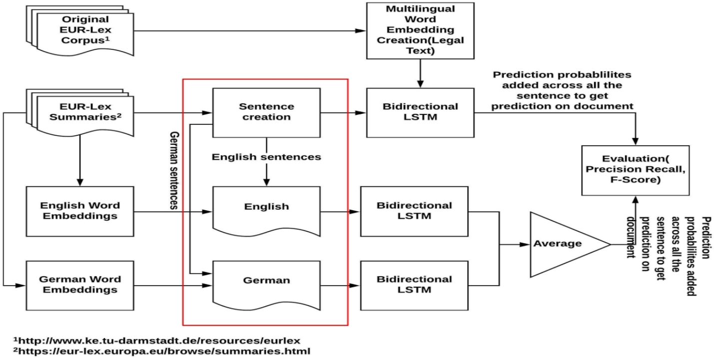


Figure: Flow chart representing the work-flow for the third research question

Approach and Implementation

- **Classifier Training**

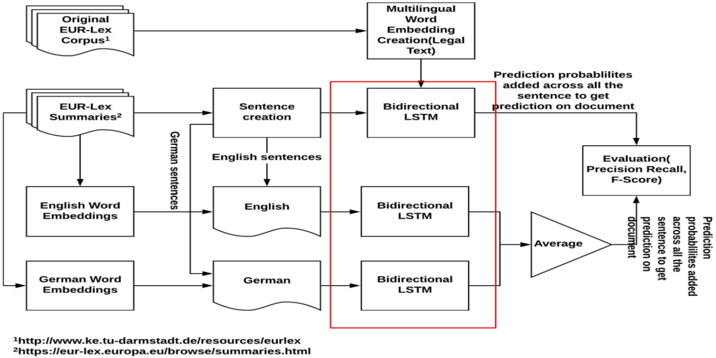


Figure: Flow chart representing the work-flow for the third research question

Approach and Implementation

- **Evaluation**

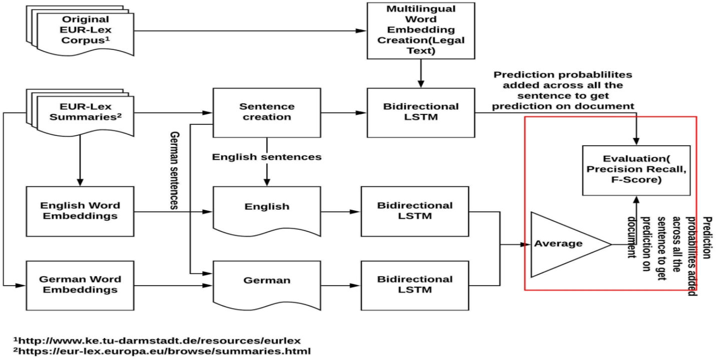


Figure: Flow chart representing the work-flow for the third research question

Approach and Implementation

Corpus

- EUR-Lex Summaries, from the Publication office of the European Union's Parliament.
- Divided into 32 topics.
- Imbalanced and Multi-labeled dataset.
- Available in multiple languages (English and German are considered here).
- Assignment of the summaries frequently changes as laws changes or are added.

Approach and Implementation

Data Preprocessing

- Removes unnecessary information from textual data

English Corpus

- Remove *stop words*.
- Lemmatization.
- Remove unnecessary symbols.
- Remove numbers and punctuations.

German Corpus

- Remove *stop words*.
- Lemmatization.
- Remove unnecessary symbols.
- Remove numbers and punctuations.
- Conversion of Umlauts to its base form.

Approach and Implementation

Sentence based approach for training BiLSTM

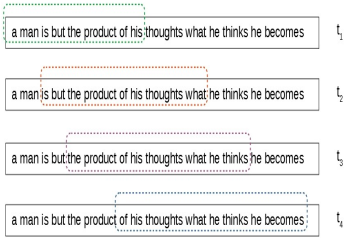


Figure: Sliding window at time step t_1, t_2, t_3, t_4 with the window size of 10 words per sentence and slide of 2 word per sentence per time step

- Every sentence of a document will have same class as document.
- Sentence length of 30 words and slide of 10 words used in creation of dataset.

Approach and Implementation

Resampling

- Exploit multi-label property of the corpus.
- Reduces samples from majority class on the basis of its presence in minority class.

Doc ID	Class Label
Document A	1,5,2
Document B	3,2,5
Document C	4,1,5
Document D	2,3
Document E	2,5

Table: An table showing documents and their assignments in their respective classes.

Class ID	No. of Samples
1	2
2	4
3	1
4	1
5	4

Table: Distribution of samples in the dataset across 5 classes.

Approach and Implementation

Resampling

Doc ID	Class Label
Document A	1
Document B	3
Document C	4
Document D	2
Document E	2 or 5

Table: Document assignment after proposed resampling

Approach and Implementation

Clustering

- To see the specialization advantage.
- Constrained K-Means clustering on the TF-IDF feature vectors of the documents.
- Elbow Analysis and Silhouettes Score to find number of clusters.
- Silhouettes Score, a measure of how similar an object is to own cluster.
- Higher Silhouettes Score, better consistency of cluster.
- Silhouettes Score for k between 2 to 8.

Approach and Implementation

Clustering

- Higher Silhouette Score and Elbow at 2 were suggestive that value of k should be 2.

k	Silhouette score
2	0.05693
3	0.04417
4	0.04749
5	0.05515
6	0.05480
7	0.05647
8	0.05537

Table: Silhouette scores for 8 values of k

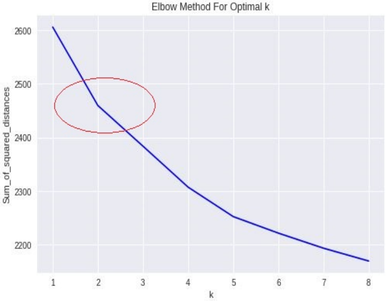


Figure: Flow chart representing the work-flow for the third research question

Approach and Implementation

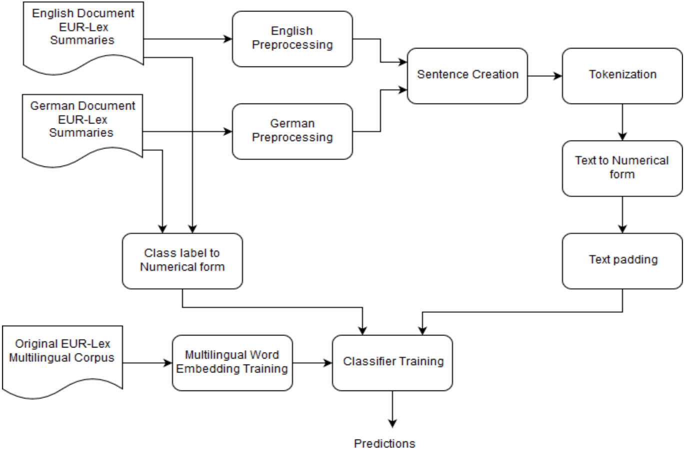


Figure: Workflow for training BiLSTM

Approach and Implementation

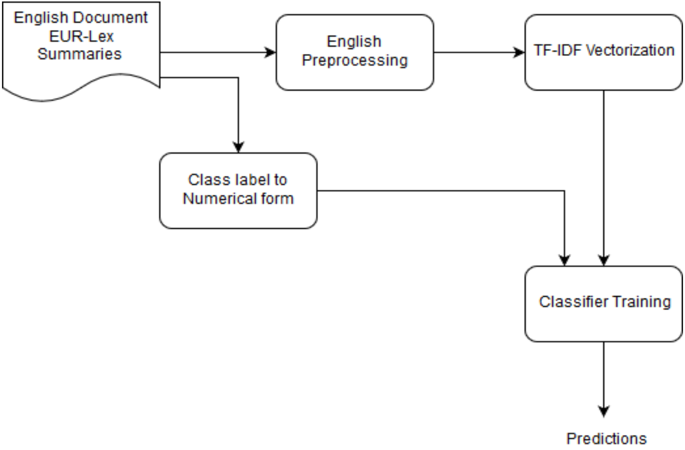


Figure: Workflow for training SVM

Results

- **Research Question 1**

Can Bidirectional Long Short Term Memory (BiLSTM) achieve better results in terms of the evaluation metrics (Precision, Recall, and F-Score) than the thoroughly studied text classification methods such as Support Vector Machines?

Results

Micro-averaged results for first research question

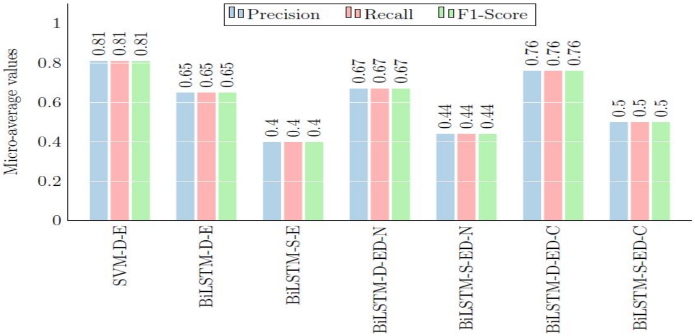


Figure: Micro-averaged *Precision*, *Recall* and *F1-Score* of SVM and BiLSTM in different configurations. The first suffix D or S indicates evaluation on document or sentence level respectively, the second suffix E or D represents the language of the corpus used respectively. The third suffix N or C indicates non clustered and clustered respectively.

Results

- **Research Question 2**

Are general-purpose resources such as pre-trained word embeddings in case of BiLSTM applicable to specific legal domain tasks in terms of evaluation metrics? Also, further training them on legal corpus, produces comparable results to the ones only trained on legal texts?

Results

Micro-averaged results for second research question

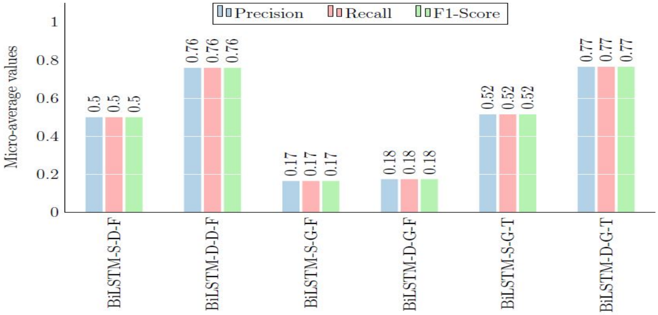


Figure: Micro-average *Precision*, *Recall* and *F1-Score* of the BiLSTM trained with general-purpose embeddings and domain-specific embeddings. The first suffix S or D indicates the evaluation on sentence or document level, the second suffix D or G represents the domain-specific or general-purpose word embeddings. The third suffix F or T indicates non trainable and trainable respectively.

Results

- **Research Question 3**

Can BiLSTM perform better when training multiple languages in a single model, compared to training one model for each language separately?

Results

Micro-averaged results for third research question

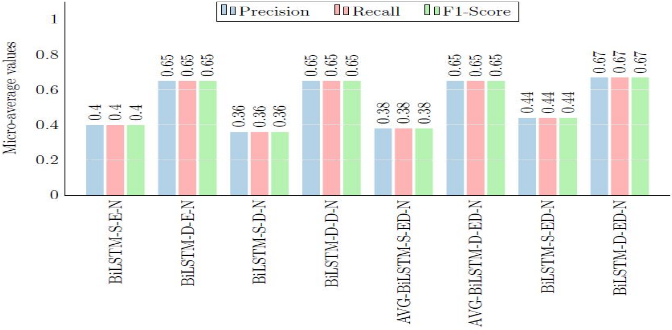


Figure: Micro-average *Precision*, *Recall* and *F1-Score* for the BiLSTM. The first suffix specifies the method of evaluation (S = Sentence and D = Document), the suffix second E, D or ED specifies English, German or both English and German respectively. The second suffix N represents that model is trained on non clustered data. AVG-BiLSTM-ED-N is the average score from BiLSTM-E-N and BiLSTM-D-N.

Conclusion

- SVMs perform better than BiLSTMs.
- Domain-specific word embedding perform better but training general-purpose embeddings yields better results.
- Adding more languages helps in improving performance of a classifier.

Conclusion

- BiLSTMs are scalable which makes it suitable when we have huge training data.
- BiLSTMs can process multiple languages.
- General-purpose embedding are trained on various sources. So training them is necessary to achieve comparable performance.
- Adding more language means having more data, which could attribute to performance increase of the classifier.

Limitation

- During data resampling, labels from majority class are removed.
- Bias towards minority classes, reduces representatives of majority classes.
- Cross-validation of classifiers.
- Clustering data for classification is counter-intuitive.

References

[1] Guido Boella, Luigi Di Caro, Llio Humphreys, Livio Robaldo, and Leon van der Torre. Nlp challenges for eunomos, a tool to build and manage legal knowledge. Language resources and evaluation (LREC),pages 3672–3678, 2012.

[2] Colas, Fabrice, and Pavel Brazdil. "Comparison of SVM and some older classification algorithms in text classification tasks." IFIP International Conference on Artificial Intelligence in Theory and Practice. Springer, Boston, MA, 2006.

[3] Corrêa Jr, Edilson A., Vanessa Queiroz Marinho, and Leandro Borges dos Santos. "Nilc-usp at semeval-2017 task 4: A multi-view ensemble for twitter sentiment analysis." arXiv preprint arXiv:1704.02263 (2017).

References

[4] Merkl, Dieter and Schweighofer, Erich The exploration of legal text corpora with hierarchical neural networks: A guided tour in public international law, 1997.

[5] Nitin Indurkhya and Fred J Damerau. Handbook of natural language processing, volume 2. CRC Press, 2010.

Classification of Legal Text Using Deep Learning: Evaluation of General-Purpose Resources for a Legal Domain-Specific Task

Master Thesis Defense

Jay Vala

Otto-von-Guericke-University Magdeburg
Faculty of Computer Sciences

June 25, 2019