NATURAL LANGUAGE PROCESSING

Assignment - 03

On the Role of Text Preprocessing In Neural Network Architectures: An Evaluation Study on Text Categorization and Sentiment Analysis

NAME – MAKSUD ALAM

ID – 16201033

COURSE – CSE431(SEC 01)

ABOUT

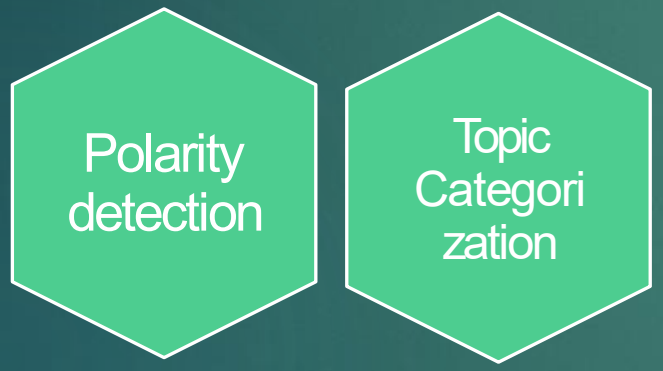
- In this paper, they emphasizes on the txt preprocessing methods for text categorization and sentiment analysis
- Here, they investigate the impact of simple text preprocessing decisions
- Provides insights into the best preprocessing practices for training word embeddings
- Focuses on the role of preprocessing of the input text and it's affects on the standard neural text classification models (like CNN)

METHODS USED FOR TEXT PREPROCESSING

- Tokenizing–given the input text, the tokenization method breaks the input text into a chunk of words.
- Lowercasing – it is the simplest preprocessing technique that convert the whole input text into lower case letter. it may negatively impact system's performance by increasing ambiguity.
- Lemmatizing– process of replacing a given token into it's corresponding lemma.
- Multiword grouping – technique that grouped consecutive tokens together into a single token.

EVALUATION

Here they used two task for the experiments :



	Dataset	Type	Labels	# of docs	Eval.
TOPIC	BBC	News	5	2,225	10-cross
	20News	News	6	18,846	Train-test
	Reuters	News	8	9,178	10-cross
	Ohsumed	Medical	23	23,166	Train-test
POLARITY	RTC	Snippets	2	438,000	Train-test
	IMDB	Reviews	2	50,000	Train-test
	PL05	Snippets	2	10,662	10-cross
	PL04	Reviews	2	2,000	10-cross
	Stanford	Phrases	2	119,783	10-cross

MODELS USED

- Here they used two classification models, one is CNN model using ReLU activation function and the second one is LSTM using softmax function.
- these models are used for both topic categorization and polarity detection

DATASETS

Topic categorization

- For the topic categorization task we used the BBCnews dataset5 (Greene and Cunningham, 2006), 20News (Lang, 1995), Reuters6 (Lewis et al., 2004) and Ohsumed7. PL04 (Pang and Lee, 2004), PL058 (Pang and Lee, 2005),RTC9,IMDB(Maas et al., 2011)

Polarity Detection

- the Stanford sentiment dataset10 (Socher et al., 2013, SF)were considered for polarity detection.

COMPARISON BETWEEN THE TWO EXPERIMENTS

PREPROCESSING EFFECT

	Preprocessing	Topic categorization				Polarity detection				
		BBC	20News	Reuters	Ohsumed	RTC	IMDB	PL05	PL04	SF
CNN	Vanilla	94.6	89.2	93.7	35.3	**83.2**	87.5	76.3	58.7†	**91.2**
	Lowercased	94.8	**89.8**	**94.2**	**36.0**	83.0	84.2†	76.1	59.6†	91.1
	Lemmatized	95.4	89.4	94.0	35.9	83.1	86.8†	75.8†	**64.2**	**91.2**
	Multiword	**95.5**	89.6	93.4†	34.3†	**83.2**	**87.9**	**77.0**	59.1†	**91.2**
CNN+LSTM	Vanilla	**97.0**	90.7	93.1	30.8†	**84.8**	**88.9**	79.1	71.4	87.1
	Lowercased	96.4	**90.9**	93.0	**37.5**	84.0	88.3†	**79.5**	**73.3**	87.1
	Lemmatized	95.8†	90.5	**93.2**	37.1	84.4	87.7†	78.7	72.6	86.8†
	Multiword	96.2	89.8†	92.7†	29.0†	84.0	**88.9**	79.2	67.0†	**87.3**

CROSS-PREPROCESSING

	Embedding Preprocessing	Topic categorization				Polarity detection				
		BBC	20News	Reuters	Ohsumed	RTC	IMDB	PL05	PL04	SF
CNN	Vanilla	94.6	89.2	93.7	35.3	83.2	87.5†	**76.3**	58.7†	**91.2**
	Lowercased	93.9†	84.6†	**93.9**	**36.2**	83.2	85.4†	**76.3**	60.0†	91.1
	Lemmatized	94.5	88.7†	93.8	35.4	83.0	86.8†	75.6	62.5	**91.2**
	Multiword	**95.6**	**89.7**	**93.9**	35.2	**83.3**	**88.1**	75.9	**63.1**	**91.2**
CNN+LSTM	Vanilla	97.0	90.7†	**93.1**	30.8†	**84.8**	**88.9**	79.1	71.4	87.1†
	Lowercased	96.4	**91.8**	92.5†	30.2†	84.5	88.0†	79.0	**74.2**	87.4
	Lemmatized	96.6	91.5	92.5†	31.7†	83.9	86.6†	78.4†	67.7†	87.3
	Multiword	**97.3**	91.3	92.8	**33.6**	84.3	87.3†	**79.5**	71.8	**87.5**

CONCLUSION

- Their evaluations highlighted the importance of being careful in the choice of how to preprocess data and to be consistent when comparing different systems.
- Their analysis showed that there is a high variance in the results depending on the preprocessing choice.