Skip Truncation for Sentiment Analysis of Long Review Information based on Grammatical Structures

Mengtao Sun[1], Ibrahim A. Hameed[1], Hao Wang[2] and Mark Pasquine[3]

[1] Department of ICT and Natural Sciences, NTNU, 6009 Ålesund, Norway
[2] Department of Computer Science, NTNU, 2815 Gjøvik, Norway
[3] Department of International Business, NTNU, 6009 Ålesund, Norway
mengtao.sun@ntnu.no, ibib@ntnu.no, hawa@ntnu.no, mapa@ntnu.no

Abstract. In reality, some emotional utterances are especially long, such as the critiques of a movie. Compared with short sentences, the sentiment of a longer paragraph is more difficult to be detected. Moreover, longer paragraphs are eased to overflow the GPU memory and sluggish to fit a large-parameterized network. Currently, the solutions include using a sentence-level summary; if not applicable, a long review is always directly truncated to an accepted length. However, the first solution will ignore the sentiment subtleties; The second solution will lose some important emotive expressions. This work puts forward a strategy to locate the potential sentiment information considering the grammatical structures and then proposes skip truncation to shorten the long review texts. This method can effectively delete irrelevant words without hitting the sentiment skeleton. The new skip truncation is compared against several baselines in binary and multiple sentiment perception. We conduct experiments with four types of standard deep neural networks. Experimental results verify that skip truncation can help reduce sentence length and improve performance in a large margin.

Keywords: Sentiment Analysis, Deep Learning, Long Text, Grammatical Structures.

1 Introduction

Many reviews are especially long in the real-world applications of sentiment analysis [1], such as the dedicated reviews of movies, products, artworks, video games, and politics. Long paragraphs easily confuse the sentiment detection models because of too many emotionless words therein [2]. Moreover, they usually overflow the GPU memory and are sluggish to fit a deep network due to their oversize [3]. This problem can be well avoided for some NLP tasks, such as named entities recognition and machine translation because a document-level input can be divided into several sentence-level batches. However, sentiment perceptron is required to perceive the overall information before predicting. Moreover, the emotion of each sentence in a document is diverse, so using sentence-level inputs will also lead to a great error. Consequently, it is important to find a solution for long review information sentiment analysis.

Currently, the solutions are still simple and insufficient [4]. The ideal way for sentiment analysis is to use a sentence-level input, but a summary of long text hardly comprises the emotional subtleties. More general practice is to truncate long paragraphs to a proper length. However, such an operation will lose much emotive content. On the one hand, feeling expressions may appear later in the reviews. Truncation may result in the texts seldom containing sentiment features. On the other hand, the sentiment reversals frequently appear in long texts, and the direct use of text truncation makes the model ignore this information. For a sentence example, "I do not want to say that the actor did a good job", by post-truncation, "I do not want to say" makes the sentence Negative, while by the pre-truncation, "The actor did a good job" makes the sentence Positive. Different snippets will influence the overall sentiment detection, and these errors may mislead the training of the sentiment analysis model. According to the perspective of human sentiment analysis, a more accurate judgment will be obtained when the texts contain many emotional expressions.

We assume that the length of input sentences for deep models generally does not exceed 200 words A larger input size often leads to memory overflow. In order to solve the abovementioned problems, this paper proposes a skip truncation. Through skip truncation, the information irrelevant to sentiment analysis can be reduced as much as possible. The truncated sentence maintains the original sentiment skeleton. This method is based on grammatical structure. Grammatical sentiment components can be a word, phrase, or clause that functions as an adjective or adverb to provide additional sentiment information. In this way, we can roughly capture the emotions in the early training stage. To summarize, this paper makes the following contributions to the sentiment analysis on long texts:

1. This paper proposes a new text truncation method for better sentiment analysis, called skip truncation.

2. This paper compares the performance of the skip truncation against different baselines in several typical deep neural networks.

3. This paper analyzes the performance of skip truncation in binary polarity classification and multiple polarity classification.

2 Related Works

Long text processing has always been a difficulty in deep learning training. To address the large scale of words, Pappagari et al. [5] use a hierarchical strategy to fit the Transformer [6] (a SOTA model in many NLP tasks) for document-level classification. They segmented the input into smaller chunks, fed them into the Transformer, then propagated each output through a single recurrent layer. He et al. [7] pointed out that document classification requires extracting high-level features from low-level word vectors. They proposed a new model incorporating the recurrent attention learning framework, focusing its attention on the discriminative parts. Wu et al. [8] proposed the hierarchical aspect-oriented framework for long document sentiment classification. Their model alleviated the challenge of the unstable sentiment information of the target aspect in long documents and the problem that the too-long input sequence can cause the model to forget previously learned information. Kaliamoorthi et al. [9] proposed a novel projection attention neural network that combines trainable projections with attention and convolutions. Their model is particularly effective on multiple large document text classification tasks.

More useful methods are based on the dictionary to address long texts sentiment analysis. First, Zhang et al. [10] extended the sentiment dictionary from degree adverb, network word, and negative word. The sentiment value of a micro-blog text is obtained by calculating the weight from the extended sentiment dictionary and Microblog texts on a topic are classified as positive, negative, and neutral. Okango et al. [11] employed a dictionary-based method for sentiment polarization from tweets related to coronavirus posted on Twitter and analyzed the effects and response in the pandemic. Yekrangi et al. [12] explored a wide related corpus using lexical resources and proposed a hybrid approach to building a lexicon specialized for financial markets sentiment analysis. Hossen et al. [13] proposed an improved lexicon-based model aiming at movie review data.

3 Materials and Method

This work proposes the use of skip truncation to construct a sentiment skeleton in order to reduce the scale of document content and amplify the subtle sentiment features. The overall workflow can be represented as follows: Raw Data Processing → Grammatical Components Tagging → Skip Truncating → Sentiment Detecting.

3.1 Raw Data

This study experimented with the idea of skip truncation by using Norwegian languages because the Norwegian language is more challengeable on text modeling due to its rich grammar rules, and it is often ignored in current sentiment analysis research [14]. Nevertheless, we believe the method can be applied to any language.

The experimental dataset is NoReC [14] (Full name: The Norwegian Review Corpus), proposed by University of Oslo. It is a Norwegian languages review dataset comprising a range of domains, including literature, movies, video games, restaurants, music, and theater, from eight rating websites in Norway. The release of the corpus consists of more than 35,000 reviews. Each review is rated with a numerical score on a scale of 1–6 and can be used for training and evaluating models for document-level sentiment analysis. These were defined by first sorting all reviews for each category by publishing date and then reserving the first 80% for training, the subsequent 10% for development, and the final 10% for testing.

The document-level reviews comprise averagely 426 words, which is lengthy for deep learning models [8, 14]. To cope with the long contents, the attributions of each document are summarized in a JSON file, containing the document ID, sentiment rating, and the manual excerpt of a document.

Data Cleaning. NoReC has already processed most of the noisy data. Here, we remove extra delimiters such as spaces and paragraph breaks. Then, we remove punctuations but keep periods, exclamations, and questions to identify sentence ending. We use word-level segmentation. In this way, the cleaned document is transformed into a sentence and can be represented as a long vector.

3.2 Sentiment on Grammatical Components

Next, the long sentences are annotated according to the linguistic structure, which is achieved by the spaCy toolkit. We use part-of-speech (POS) and syntactic dependency relation to label each token. POS and syntactic dependency relation can be regarded as the tags based on the word-level and the sentence-level grammars. We investigate each component that may be closely related to sentiment classification. Finally, according to empirical observations, the importance is summarized in Table 1 and Table 2.

Table 1. Grammatical sentiment components based on POS tagging.

POS	Description	POS	Description
ADJ	Adjective	INTJ	Interjection
ADV	Adverb	AUX	Auxiliary
PART	Particle	PUNCT	Punctuation
VERB	Verb	O	All other tags in UPOS[1].

As a result, we divide each word into three categories, red, blue, and non-sentiment. Red marks will greatly impact sentiment classification; Blue marks are secondary grammatical sentiment components.

ADJ and ADV are the main research objects, such as ‘det **store** huset’ (the **big** house), ‘**nesten** ferdig’ (**almost** finished). Some studies utilize ADJ and ADV to compute sentiment scores and construct sentiment dictionaries [10]. The words in PART tag cause an emotional reversal, such as ‘Han liker **ikke** å spise is’ (He **does not** like to eat ice cream). INTJ tag is mostly a phrase used for exclamations, such as ‘**bravo**’. AUX can enhance the expressive intensity, such as ‘**burde**’ (should), ‘**må**’ (must), etc.

Some words associated with PUNCT tags can enhance emotional expression. Moreover, some verbs have strong emotional tendencies, such as ‘**liker**’ (like), but the VERB tag comprises a wide scope. Considering the actual situation of a document, the number of emotional expressing is limited. PUNCT and VERB are attributed to the secondary level. An elaborate specification of POS tags can be found in UPOS[1].

In sentence-level grammar, syntactic dependency relations describe the reliance between words. It points out the collocation between words in terms of their semantics. The auto-annotation has achieved 89% accuracy in the best Norwegian model. Simi-

[1] https://universaldependencies.org/u/pos/

lar to POS tagging, based on the syntactic dependencies, the importance of sentiment classification is depicted in Table 2.

Table 2. Grammatical sentiment components based on syntactic dependency tagging.

Dependency	Description	Dependency	description
advmod	adverbial modifier	Root	root
amod	adjectival modifier	Punct	punctuation
aux	auxiliary	O	all other tags in UD[2]

Syntactic dependency relations are more complicated, but the labeling of grammatical sentiment components is equivalent to Table 1. For example, advmod is an adverb or adverbial phrase that modifies a predicate or modifier word, parallel to ADV in Table 1. An elaborate specification of syntactic dependency relations can be found in Universal Dependency[2].

3.3 Skip Truncation

Since the punctuations of ending the sentences are retained, we iteratively pass sentence-level input to obtain POS tags and syntactic dependency relations for every word in a document. After receiving all the grammars, we delete the non-sentiment words both in POS and syntactic dependency standard (See Table 1 and Table 2). Then, if the sequence length exceeds the threshold (200 in this work), we execute the following command and stop when the document length is acceptable:

1. Delete the blue words that appear once in POS or dependency,
2. Delete the blue words that appear both in POS and dependency,
3. Delete the red words that appear once in POS or dependency.

In this way, the document has been turned into sporadic words. Although the sequence is broken, the sentiment skeleton has been preserved. For documents whose length is 700, the truncated length is usually less than 100. If the file size is still more than the threshold, directly truncate the part exceeding the threshold. If the sentence length is less than the threshold, use zero padding after the sequence.

3.4 Sentiment Perceptron

We introduce a concise architecture for sentiment classification and explore how the neural networks behave in different truncations. The specification is shown as follows:

Embedding. In skip truncation, the context information is broken. We experimented with a pre-trained embedding and found that the pre-trained embedding can hardly get a satisfactory result, so it is suggested to train each model from scratch. To reduce overfitting, we apply dropout in the embedding layer.

[2] https://universaldependencies.org/u/dep/

Hidden Neurons. After embedding, we verify four types of neurons for extracting sentiment features: Fully Connection units, convolutional units, Recurrent units, and Gated Recurrent units [15].

Classification. The classification part contains two layers of fully connection. The first layer furtherly extracts and integrates features, and the second layer is used for sentiment labeling.

4 Experimental Results

In experiments, we test ship truncation on binary sentiment polarity and multiple (six) sentiment polarity by different hidden neurons settings. This section first introduces the model settings, then illustrates the baselines, and finally presents the experimental results.

4.1 Model Settings

According to the architectures of sentiment perceptron, the details of parameters are introduced in Table 3. The model comprises one of the Hidden Neurons and one of the Classification settings each time, while the Embedding and Training settings are the same.

Table 3. Parameters of sentiment models.

Component	Value	Explanation
Embedding	sentence length 200, embedding dimension 300, dropout rate 50%	Embeddings are trained from scratch

Hidden Neurons	Fully-Connection: unit 50
	1D Convolution: unit 50, kernel size 3
	bidirectional Recurrent: 50 units
	bidirectional Gated Recurrent: 50 units
		for Fully-Connection Unit
		for Convolutional Unit
		for Recurrent Unit
		for Gated Recurrent Unit

Classification	2 Fully-Connection layers: unit 32 and unit 6
	2 Fully-Connection layers: unit 32 and unit 2
		for 6 sentiment polarities
		for 2 sentiment polarities

Training	Optimizer: Adam, Activation Function: ELU, Learning Rate: 0.001, Batch Size: 64, Epoch: 5	-

To evaluate the performance of our proposed model, we applied standard Precision (P), Recall (R), F1 Score between the model output and ground truth.

4.2 Baselines

We use three baseline truncations to prove the performance of the skip truncation. Truncated sentences have an identical fixed sentence length, 200, and all the experiments are performed in the same model setting.

Excerpt. Documents are manually summarized with sentence-level excerpts, and each excerpt is supplemented to 200 dimension using zero-padding in the end.

Brute Force I. Each document is selected 200 words from the beginning, using zero-padding to the end if the entire document is less than 200 words.

Brute Force II. Each document is selected 200 words from the end, using zero-padding to the beginning if the entire document is less than 200 words.

4.3 Results on Binary Sentiment Polarity

Binary sentiment polarity classification is more manageable due to its lower requirements on featurization. Therefore, we first integrate the original six sentiment labels to negative and positive. Labels 1, 2, 3 are negative and labels 4, 5, and 6 are positive. The performances of different neural networks against baselines are shown in Table 5-8.

Table 5. Results in fully connection (Binary-Polarity).

Fully Connection	Precision	Recall	F1 Score
Excerpt	69.60	74.77	71.12
Brute Force I	68.50	75.80	66.29
Brute Force II	76.00	79.24	72.72
Skip-Truncation	80.56	80.50	76.65

By fully connection, the F1 score of skip truncation reaches 76.65%, the Precision reaches 80.56%, and Recall reaches a rate of 80.50%. The F1 score, Precision, and Recall, is higher than the second about 4%, 4%, and 1%, respectively. On the other hand, the fully connection is not strong enough for featurization, which cannot display the full advantages of skip truncation.

Table 6. Results in convolutional unit (Binary-Polarity).

Convolutional Unit	Precision	Recall	F1 Score
Excerpt	69.60	75.57	70.54
Brute Force I	77.93	79.13	72.31
Brute Force II	83.68	84.86	82.92
Skip-Truncation	90.80	90.94	90.58

Convolutional units can better consider the relationship of neighbor words. The F1 score, Precision and Recall reaches 90.58%, 90.94% and 90.94%. Compared to corresponding metrics 83.68%, 84.86%, 82.92% in second place, the skip truncation shows stronger effectiveness.

Table 7. Results in recurrent unit (Binary-Polarity).

Recurrent Unit	Precision	Recall	F1 Score
Excerpt	69.14	73.51	70.83
Brute Force I	68.30	71.56	69.54
Brute Force II	69.21	68.12	68.63
Skip-Truncation	93.13	93.23	93.15

The bidirectional recurrent architecture is especially suitable for text modeling because it considers the context from each time step. The performances by skip truncation manifest a further enhancement and achieve 93.15% F1 score, 93.13% Precision and 93.23% Recall.

Table 8. Results in gated recurrent unit (Binary-Polarity).

Gated Recurrent Unit	Precision	Recall	F1 Score
Excerpt	70.39	73.05	71.52
Brute Force I	74.01	75.69	74.71
Brute Force II	80.51	81.65	80.93
Skip-Truncation	93.83	93.92	93.83

The Gated recurrent unit is an improved version of the recurrent unit. The skip truncation keeps an equivalent good performance in a gated recurrent network. It produces a 93.83% F1 score, 93.83% Precision, and 93.92% Recall.

Overall, the Brute Force II method is better than Brute Force I because reviews include non-subjective descriptions in the front of the document, such as introducing the movie characters. Compared with baselines, skip truncation can improve the accuracy of binary sentiment polarity. It is especially effective for those high-capable networks. It can break through the limitation of sentiment featurization and the most correctly perceive the document sentiment.

4.4 Results on Multiple Sentiment Polarity

In multiple sentiment polarity, the sentiment gradually changes from negative to positive from level one (negative) to level six (positive). The difficulty is perceiving the subtle sensation. The performances against baselines are shown in Table 9-12.

Table 9. Main result in fully connection (Multiple-Polarity).

Fully Connection	Precision	Recall	F1 Score
Excerpt	33.01	37.61	33.66
Brute Force I	31.42	36.47	32.02
Brute Force II	35.07	39.91	35.83
Skip-Truncation	51.77	50.92	47.57

Table 10. Main result in convolutional unit (Multiple-Polarity).

Convolutional Unit	Precision	Recall	F1 Score
Excerpt	34.27	35.32	34.55
Brute Force I	39.66	38.07	33.71
Brute Force II	47.87	49.54	46.53
Skip-Truncation	64.85	64.91	63.44

Table 11. Results in recurrent unit (Multiple-Polarity).

Recurrent Unit	Precision	Recall	F1 Score
Excerpt	31.58	34.17	32.66
Brute Force I	28.49	29.93	29.10
Brute Force II	36.11	36.12	36.00
Skip-Truncation	85.17	84.98	84.89

Table 12. Results in gated recurrent unit (Multiple-Polarity).

Gated Recurrent Unit	Precision	Recall	F1 Score
Excerpt	35.78	39.56	36.93
Brute Force I	39.85	39.56	38.51
Brute Force II	50.11	50.92	49.78
Skip-Truncation	84.86	84.63	84.31

Experiments showed more significant differences in the multiple sentiment polarity because detailed expression changing will influence the prediction and the adjacent sentiment label is very hard to be discriminated. Merely by different truncation directions (see Brute Force I and II), the results showed a more than 10% gap in Precision, Recall, and F1 score. However, the baselines cannot sufficiently capture the fine-grained feelings. The best baseline got 50.11% Precision, 50.92% Recall, and 49.78% F1 score, while skip truncation improved them to 85.17%, 84.98%, and 84.89%. The results manifested that skip truncation can keep the correct sentiment features in a document and enable models to receive the complete sentiment skeleton. It can effectively reduce the interference of irrelevant words and amplify subtle information.

5 Conclusion

The document-level sentiment classification is difficult for deep learning training [1-8], while current truncations will reduce the expressive features and thus hinder the model capability. This paper proposed a new skip truncation method based on grammatical structures. The proposed truncation can reduce document length and enhance subtle sentiment detection. Using skip truncation for binary and multiple polarity got the best performances in different neural networks against baselines.

In future work, we will study more grammatical components towards sentiment information to keep the features extraction when truncation. We will also discuss more applications using the skip-truncation approach for text modeling.

References

1. Yin, Y., Song, Y. and Zhang, M.: Document-level multi-aspect sentiment classification as machine comprehension. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 2044-2054. ACL, Copenhagen, Denmark (2017).
2. Rhanoui, M., Mikram, M., Yousfi, S. and Barzali, S.: A CNN-BiLSTM model for document-level sentiment analysis. Machine Learning and Knowledge Extraction, 1(3), 832-847 (2019).
3. Adhikari, A., Ram, A., Tang, R., Hamilton, W.L. and Lin, J.: Exploring the limits of simple learners in knowledge distillation for document classification with DocBERT. In: Proceedings of the 5th Workshop on Representation Learning for NLP, pp. 72-77. ACL, Online Meeting (2020).
4. Fiok, K., Karwowski, W., Gutierrez, E., Davahli, M.R., Wilamowski, M. and Ahram, T.: Revisiting Text Guide, a Truncation Method for Long Text Classification. Applied Sciences, 11(18), 8554 (2021).
5. Pappagari, R., Zelasko, P., Villalba, J., Carmiel, Y. and Dehak, N.: Hierarchical transformers for long document classification. In: 2019 IEEE Automatic Speech Recognition and Understanding Workshop, pp. 838-844. IEEE, Singapore (2019).
6. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems, pp. 5998-6008. Curran Associates, Long Beach, CA, USA (2017).
7. He, J., Wang, L., Liu, L., Feng, J. and Wu, H.: Long document classification from local word glimpses via recurrent attention learning. IEEE Access, 7, 40707-40718 (2019).
8. Wu, Z., Gao, J., Li, Q., Guan, Z. and Chen, Z.: Make aspect-based sentiment classification go further: step into the long-document-level. Applied Intelligence, 1-20 (2021).
9. Kaliamoorthi, P., Ravi, S. and Kozareva, Z.: PRADO: Projection attention networks for document classification on-device. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pp. 5012-5021. ACL, Hong Kong, China (2019).
10. Zhang, S., Wei, Z., Wang, Y. and Liao, T.: Sentiment analysis of Chinese micro-blog text based on extended sentiment dictionary. Future Generation Computer Systems, 81, 395-403 (2018).
11. Okango, E. and Mwambi, H.: Dictionary Based Global Twitter Sentiment Analysis of Coronavirus (COVID-19) Effects and Response. Annals of Data Science, 1-12 (2022).
12. Yekrangi, M. and Abdolvand, N.: Financial markets sentiment analysis: Developing a specialized Lexicon. Journal of Intelligent Information Systems, 57(1), 127-146 (2021).
13. Hossen, M.S. and Dev, N.R.: An Improved Lexicon Based Model for Efficient Sentiment Analysis on Movie Review Data. Wireless Personal Communications, 1-10 (2021).
14. Velldal, E., Øvrelid, L., Bergem, E.A., Stadsnes, C., Touileb, S. and Jørgensen, F.: NoReC: The norwegian review corpus. In: Proceedings of the 11th edition of the Language Resources and Evaluation Conference, pp. 1–2. ACL, Miyazaki, Japan (2018)
15. Zhang, L., Wang, S. and Liu, B.: Deep learning for sentiment analysis: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8(4), e1253 (2018).