Sentence-Level Sentiment Classification A Comparative Study Between Deep Learning Models

Sara Mifrah* and El Habib Benlahmar

Laboratory of Information Processing and Modelling, Hassan II University of Casablanca, Faculty of Sciences Ben M'sik, Casablanca, Morocco
E-mail: mifrah.sara@gmail.com
**Corresponding Author*

Received 22 February 2022; Accepted 08 March 2022;
Publication 11 May 2022

Abstract

Sentiment classification provides a means of analysing the subjective information in the text and subsequently extracting the opinion. Sentiment analysis is the method by which people extract information from their opinions, judgments and emotions about entities. In this paper we propose a comparative study between the most deep learning models used in the field of sentiment analysis; L-NFS (Linguistique Neuro Fuzzy System), GRU (Gated Recurrent Unit), BiGRU (Bidirectional Gated Recurrent Unit), LSTM (Long Short-Term Memory), BiLSTM (Bidirectional Long Short-Term Memory) and BERT(Bidirectional Encoder Representation from Transformers), we used for this study a large Corpus contain 1.6 Million tweets, as devices we train our models with GPU (graphics processing unit) processor. As result we obtain the best Accuracy and F1-Score respectively 87.36% and 0.87 for the BERT Model.

Keywords: Sentiment classification, sentence level, deep learning, BiLSTM, LSTM, GRU, BiGRU, L-NFS, transformer, BERT.

Journal of ICT Standardization, Vol. 10_2, 339–352.
doi: 10.13052/jicts2245-800X.10213

1 Introduction

Natural language processing, a branch of machine learning, has become immensely popular in the past 6 years in both academic research and industrial applications due to advances in deep learning and increased computing power of hardware systems. It is a technique that allows computers to comprehend the functioning of human expression by making use of computational linguistics and the field of computer science. In addition, NLP has been used in recent years in several applications related to the understanding and interpretation of text, audio and video files.

Deep Learning is a field of artificial intelligence, and more specifically a subfield of Machine Learning. It is based on letting machines learn from their experiences, as humans do. The main difference with Machine Learning is that Deep Learning algorithms do not truly have a limit in their training capacity. The more they are given data to train on, the more these systems improve their efficiency.

The recurrent neural network (RNN) is a type of artificial neural network (ANN) used for time-series retrieval. One of the attractions of RNNs [1] is the notion that they could be able to relate previously acquired information to the current task, such as using previous video images to inform comprehension of the present image. If RNNs were able to do this, they might be incredibly useful. However, even looking at recent information is sometimes all that is needed to accomplish the current task. Take the example of a language model that tries to predict the next word based on previous words. If we are trying to predict the last word of "the sky is full of clouds", we don't need any additional context – it is pretty obvious that the next word will be "clouds". In cases like this, where the distance between the relevant information and where it is needed is low, RNNs can be taught to use past information [1, 2]. But there are also instances where we need more context.

A transformer represents a deep learning model that takes the self-attention mechanism, by weighting the relevance of each part of the input data in a differentiated way. It is primarily used in natural language processing (NLP) [3]. Transformers were first introduced in 2017 by a team at Google Brain [3] and are increasingly the model of choice for NLP problems [5], displacing RNN models such as Long Short Term Memory (LSTM). The further training parallelization allows training on bigger datasets than was once possible. This led to the development of pre-trained systems such as BERT [19] (Bidirectional Encoder Representations from Transformers), which have been trained with large linguistic datasets, such as the Wikipedia Corpus, and can be refined for specific tasks [4, 7].

Sentiment analysis currently has emerged as the dominant approach used to extract feelings and opinions from online sources. Sentiment analysis concentrates on splitting linguistic units into two classes: objective and subjective, while sentiment analysis seeks to divide language entities into three categories: negative, positive and neutral.

This paper is structured as follows: Section 2 presents some related works. Section 3 presents the methodology, our experimentations and results are described in Section 4, and finally the conclusions of this paper are provided in Section 5.

2 Related Work

We will briefly review related works on sentence-level sentiment classification and neural networks for sentiment classification.

2.1 Sentiment Classification at Sentence-Level

Sentence-level sentiment classification is typically considered as a classification problem in the literature. In the classification of sentiment at the sentence level, a sentence is classified as either subjective or objective. A subjective statement indicates an opinion about an entity. For example, "I have a nice sack", means a positive polarity about the sack. Therefore, it is considered a subjective statement that can be classified into various polarities. On the other hand, factual statements are called objective statements. A statement such as "The colour of the bottle is blue," does not display any sentiment, so it is classified as an objective statement. [8] proposed a framework called Weakly Supervised Deep Embedding (WDE), which uses review scores to train a sentiment classifier. They used CNN to construct WDE-CNN and LSTM to construct WDE-LSTM to extract feature vectors from review sentences. The model was evaluated on the Amazon dataset of three domains: digital cameras, cell phones and laptops. The accuracy obtained on the WDE-CNN model was 87.7% and on the WDE-LSTM model was 87.9%, showing that deep learning models yield the highest accuracy compared to baseline models. [9] developed a model called Multi-level Sentiment-enriched Word Embedding (MSWE), which uses a multilayer perceptron (MLP) to model word-level sentiment information and CNN to model tweet-level sentiment information. The model additionally learns the embeddings of sentiment-specific words, and SVM is used for classification of sentiment. It was evaluated on the SemEval2013 dataset and the context-sensitive Twitter dataset, which are the reference datasets for the sentiment

classification tasks. The F1 score achieved on the SemEval2013 dataset was 85.75 while on the CST dataset was 81.34.

2.2 Sentiment Classification with Neural Networks

Currently, deep learning approaches have become a popular solution for treating sentiment analysis tasks. [10] The authors surveyed current work on deep learning for sentiment analysis and grouped deep learning models into recursive, non-recursive, and a mixture of both models. The authors additionally paralleled document-level and sentence-level sentiment analysis on two datasets. They conclude that deep learning models can be a better solution for polarity detection. [11] proposed a supervised learning model based on the long-term memory algorithm (LSTM). In this model, a representation of the sentence was done by catching the link of every target word with its contexts. They indicated that incorporating the target information improves the classification accuracy of the suggested model. Other researchers suggested use of deep convolutional neural networks (CNN) to extract features from multimodal content and feed these features to a multiple kernel learning classifier for sentiment detection [12]. Their method allows them to achieve good results using different datasets. Regarding to [13] they proposed a hybrid approach that utilises CNN to learn embedded features and then used a multiobjective genetic algorithm based optimization technique to generate the sentiment augmented optimised vector. They trained a Support vector machine (SVM) with a non-linear kernel for sentiment detection. In this study [17], the researchers propose a deep learning based BiLSTM technique to categorise the reviews provided by the restaurant visitors into positive and negative polarities. A Corpus consisting of 8435 reviews is composed to evaluate the proposed technique. In a subsequent step, a comparative analysis of the proposed technique with other machine learning algorithms is provided. On the testing dataset, the evaluation results show that the BiLSTM technique achieved the highest accuracy of 91.35%. [6] In this study authors compare CNN and RNN models using a corpus of citation sentiment analysis consisting of 8736 citation sentences, the results show that the BiLSTM model achieved the best accuracy of 96.26%. In addition, mixing multiple deep learning algorithms could be useful to improve the performance of sentiment analysis models. [14, 16] A deep learning model was proposed by mixing CNN and LSTM approaches and using word embedding to describe evaluation sentences. [14] This model, used LSTM as a pooling layer to capture long-term dependencies, which is one of the limitations of

the CNN algorithm. They interpreted their model using the database (IMDB) and Stanford's sentiment Treebank datasets. The proposed model solved the polarity detection task and mitigated the word order problem. Similarly, [15] combined CNN and LSTM to successfully solve the aspect-level sentiment analysis problem for news articles. This combination is useful for capturing the semantics of words and the combination between them. Another work [18] proposes a mixture of convolutional neural networks (CNN) and bidirectional long-term memory models (BiLSTM), with Doc2vec integration, suitable for opinion analysis in long texts. The CNN-BiLSTM model is contrasted with the CNN, LSTM, BiLSTM and CNN-LSTM models with Word2vec/Doc2vec incorporation. Some combined the Doc2vec model with CNN-BiLSTM were applied to French newspaper articles and the result outperformed other models with an accuracy of 90.66%.

2.3 Transformers BERT

Similar to recurrent neural networks (RNNs), transformers are made to handle sequential data, like natural language, for such tasks as translation and text summarization. In contrast to RNNs, however, transformers do not demand that sequential data be parsed in order. If the input data is a natural language sentence, for example, the transformer need not process the beginning before the end. With this feature, the transformer allows for much greater parallelization than RNNs and thus reduced training times. Transformers have quickly become the model of choice for NLP problems, replacing older recurrent neural network models such as LSTM (Long Short-Term Memory). Because the transformer model facilitates more parallelization during training, it has allowed training on larger datasets than was possible before its introduction. This has led to the development of pre-trained systems such as BERT (Bidirectional Encoder Representations from Transformers), which have been trained with huge general text datasets, such as Wikipedia Corpus, and can be refined to perform specific linguistic tasks.

3 Methodology

3.1 Data Set

In this study, we choose to use the Sentiment140 Twitter datasets to learn and predict the sentiments of 1,600,000 tweets extracted using the twitter api. The tweets have been annotated (0 = negative, 2 = neutral, 4 = positive). The Sentiment140 dataset, unlike others, is a semi-structured dataset where

although it contains 6 columns (Sentiment, id, Data, Query, User, Tweet), the only fields that are useful for our analysis are primarily Tweet and Sentiment. Hence the dimensionality of this dataset cannot be easily determined. This raw data comprises sentences which range between 1 and 250 characters resulting in an average of 30 word tokens.

Table 1 Example of Dataset Structure with the two primarily data (Tweet and Sentiment)

	Sentiment	Tweet
0	0	@switchfoot http://twitpic.com/2y1zl – Awww, t...
1	0	is upset that he can't update his Facebook by ...
2	0	@Kenichan I dived many times for the ball. Man...
3	0	my whole body feels itchy and like its on fire
4	0	@nationwideclass no, it's not behaving at all...
5	0	@Kwesidei not the whole crew
6	0	Need a hug

- ***Sentiment distribution***

the graph in below presents the distribution of sentiment in dataset after renaming sentiment (0,4) with respectively (Negative, Positive) (Figure 1):

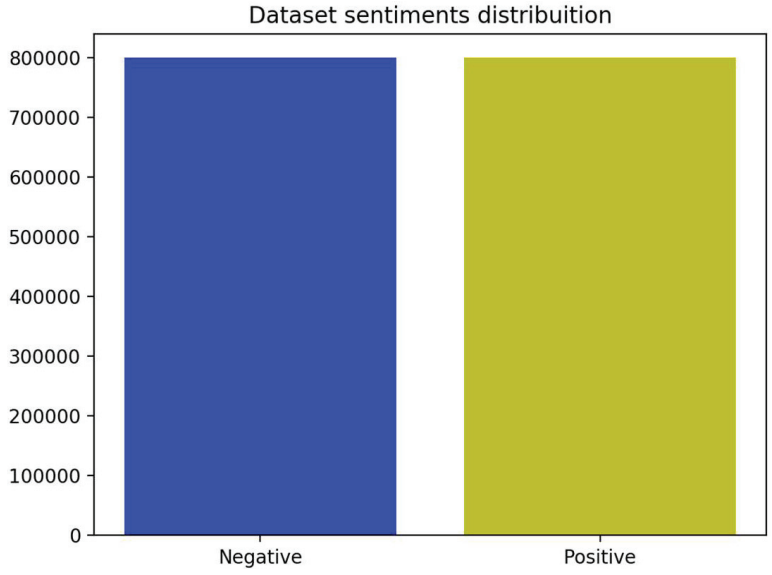

Figure 1 Sentiment distribution in dataset between (Negative, Positive) respectively (0,4).

- ***Word Cloud***

Figure 2 Dataset word cloud.

3.2 Pre-processing

In Natural Language Processing (NLP) and Information Retrieval (IR) the step of Preprocessing is a crucial step in making the text more digestible by removing nonsense phrases, noise, and unnecessary repetitions so that a deep model can boost their efficiency. Perversely, the language used on social media is nonstandard and informal, and noise requires extensive processing before feeding the network. The following steps are performed as part of data preparation:

Firstly we load the dataset, we normalise target labels we need map value 4 as 1 (i.e. Positive) and 0 as 0 (i.e. Negative), then remove duplicates identified by a combination of user, tweet and sentiment label and remove special and non ASCII characters, we replace contractions with complete phrases (e.g.: "I'm" with "I am"), after that we expand all abbreviations and other urban lingos, we also remove stopwords by Using NLTK library's

English stopword dictionary and finely we remove email addresses, URLs, twitter user handles, etc. After the data is cleaned we remove any records that end up with a zero-length tweet.

Then we switch to the main three tasks for the Pre-processing phase :

- *Tokenization is the method of dividing a text into words, phrases or other meaningful parts, namely tokens. In other words, tokenization is a type of text segmentation. Generally, segmentation is performed by considering only alphabetic or alphanumeric characters that are delimited by non-alphanumeric characters (e.g. punctuation, white spaces).*
- *Stemming goal is, to obtain the stem forms, or roots, of derived words. Because derived words are semantically similar to their root form, word occurrences are usually calculated after applying stemming to a given text.*
- *Conversion to lower case is another widely used preprocessing step. Since there is supposed to be no difference between upper and lower case forms of words, all upper case characters are usually converted to their lower case forms.*

4 Experimentations and Results

After the preprocessing phase we tried to execute and compare our five models; LSTM, Bi-LSTM, GRU, L-NFS and BERT, then calculated their accuracy, F1-Score and Loss Function and finally analysed the results obtained.

4.1 Parametres

In this study we use a large corpus with 1 600 000 tweets, We split the dataset between the train part with 75%, validation part with 5% and the rest for testing part (Table 2):

Table 2 Values of each split part: training, testing and validation

TRAIN size	1200000
TEST size	320000
VALIDATION size	80000

then we use different parameters for each model, for example we use GPU for training BERT model and CPU for others models, as well as for Optimiser we use AdamW for BERT model and Adam for others RNN models.

Table 3 Some Parameters using for training models

Activation Function	Sigmoid/Softmax
Optimiser	Adam/AdamW
Device	CPU or GPU

4.2 Results

Table 4 concludes the accuracy, Loss Function and F1-Score of different models L-NFS, GRU, BiGRU, LSTM, BiLSTM and BERT performances on the Sentiment140 dataset with some parameters used in the training phase.

Accuracy: BERT **87.36%**, BiLSTM **79.73%**, BiGRU **79.02%**, LSTM **78.64%**, L-NFS **62.49%**, GRU **50.03%**

Loss Function: BERT **0.30**, BiLSTM **0.41**, BiGRU **0.43**, LSTM **0.43**, L-NFS **0.48**, GRU **0.49**

4.3 Discussion

The objective of this study is to present an overview and comparison of the most widely used deep learning models applied to a large corpus for sentence-level sentiment classification.

We observe that the transformer BERT is able to classify sentiment polarity with an accuracy of 87.4%. For LSTM and BiLSTM their accuracy achieved respectively 78.6% and 79.7% with a Dropout 0.2, Batch-size = 1204, Epoch = 10, and we use as Word Embedding the GloVe Embedding. BiGRU achieved 79% with 32 Batch-size and 4 Epoch, L-NFS with 62.5%. The last one is GRU has the minimum accuracy 50% with 32 Batch-size and 70 Epoch.

Table 4 Results (Accuracy, Loss Function, F1-Score) after treatment for each model with some parameters (Dropout, Batch_size and Epoch)

Model	Accuracy	Loss Function	F1-Score	Dropout	Batch_size	Epoch
L-NFS	*0.6249*	*0.4780*	*0.70*	–	–	–
GRU	*0.5003*	*0.4971*	*0.50*	*0.2*	*32*	*70*
BiGRU	*0.7902*	*0.4313*	*0.79*			*4*
LSTM	*0.7864*	*0.4327*	*0.79*		*1024*	*10*
BiLSTM	*0.7973*	*0.4120*	*0.80*			
BERT	***0.8736***	***0.2971***	***0.87***	**–**	*32*	*1*

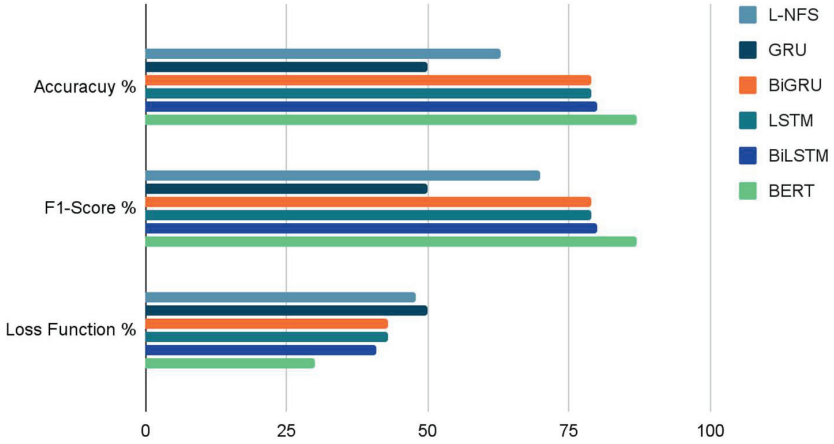

Figure 3 Models comparative graph.

5 Conclusion

Our current century is marked by significant developments in the field of artificial intelligence, and the deep learning revolution has changed the whole branch of artificial intelligence. Deep learning techniques have become essential tools of any model in today's computer world. However, deep learning techniques offer a high degree of automation with generalised rule extraction for text classification and sentiment analysis tasks.

References

[1] Pamina, J. and Raja, Beschi (2019). Survey on Deep Learning Algorithms (January 12, 2019). International Journal of Emerging Technology and Innovative Engineering, Volume 5, Issue 1, January 2019, Available at SSRN: https://ssrn.com/abstract=3351289.

[2] Alex Sherstinsky (2020). Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network, Physica D: Nonlinear Phenomena, Volume 404, 2020, 132306, ISSN 0167-2789, https://doi.org/10.1016/j.physd.2019.132306.

[3] Vaswani, Ashish; Shazeer, Noam; Parmar, Niki; Uszkoreit, Jakob; Jones, Llion; Gomez, Aidan N.; Kaiser, Lukasz; Polosukhin, Illia (2017-06-12). "Attention Is All You Need". arXiv:1706.03762.

[4] "Open Sourcing BERT: State-of-the-Art Pre-training for Natural Language Processing". Google AI Blog. Retrieved 2019-08-25.

[5] Wolf, Thomas; Debut, Lysandre; Sanh, Victor; Chaumond, Julien; Delangue, Clement; Moi, Anthony; Cistac, Pierric; Rault, Tim; Louf, Remi; Funtowicz, Morgan; Davison, Joe; Shleifer, Sam; von Platen, Patrick; Ma, Clara; Jernite, Yacine; Plu, Julien; Xu, Canwen; Le Scao, Teven; Gugger, Sylvain; Drame, Mariama; Lhoest, Quentin; Rush, Alexander (2020). "Transformers: State-of-the-Art Natural Language Processing". Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 38–45. doi: 10.18653/v1/2020.emnlp-demos.6. S2CID 208117506.

[6] Sara Mifrah, El Habib Benlahmar, Mohamed Ezeouati, Youssef Mifrah (2022). Citation Sentiment Analysis with Deep Learning: a Comparative Study between RNN & CNN Models,The 2nd International Workshop on New Services and Networks WNSN'22.

[7] "Better Language Models and Their Implications". OpenAI. 2019-02-14. Retrieved 2019-08-25.

[8] Zhao, Wei, Guan, Ziyu, Chen, Long, He, Xiaofei, Cai, Deng, Wang, Beidou, Wang, Quan. (2017). Weakly-Supervised Deep Embedding for Product Review Sentiment Analysis. IEEE Transactions on Knowledge and Data Engineering. pp. 1–1. 10.1109/TKDE.2017.2756658.

[9] Xiong, Shufeng, Lv, Hailian, Zhao, Weiting, Ji, Donghong. (2017). Towards Twitter Sentiment Classification by Multi-Level Sentiment-Enriched Word Embeddings. Neurocomputing. 275. 10.1016/j.neucom .2017.11.023.

[10] Rojas-Barahona LM. Deep learning for sentiment analysis language and linguistics. Compass 10:701–719. https://doi.org/10.1111/lnc3.12228 (2016).

[11] Tang D, Qin B, Feng X, Liu T. Effective LSTMs for target-dependent sentiment classification. In: Paper presented at The 26th international conference on computational linguistics (COLING 2016). Osaka, Japan., 11–16 Dec. 2016, pp. 3298–3307. (2016).

[12] Poria S, Chaturvedi I, Cambria E, Hussain A. Convolutional MKL based multimodal emotion recognition and sentiment analysis. In: Paper presented at the 2016 IEEE 16th international conference on data mining (ICDM), Barcelona, Spain, pp. 439–448 (2016).

[13] Akhtar MS,Kumar A, EkbalA,Bhattacharyya P. Ahybrid. deep learning architecture for sentiment analysis. In: Paper presented at the 26th international conference on computational linguistics: technical papers, COLING, Osaka, Japan, 11–17 Dec 2016, pp. 482–493 (2016).

[14] Hassan A, Mahmood A. Deep learning approach for sentiment analysis of short texts. In: 2017 3rd international conference on control, automation and robotics (ICCAR), 24–26 April 2017, pp. 705–707. https://doi.org/10.1109/ICCAR.2017.7942788 (2017).

[15] Nguyen D, Vo K, Pham D, Nguyen M, Quan T (2017). A deep architecture for sentiment analysis of news articles. In: Le N-T, van Do T, Nguyen NT, Thi HAL (eds) Advanced computational methods for knowledge engineering: proceedings of the 5th international conference on computer science, applied mathematics and applications, ICCSAMA 2017. Springer, Cham, pp. 129–140. https://doi.org/10.1007/978-3-319-61911-8_12.

[16] Muppidi, Satish, Gorripati, Satya Keerthi, and Kishore, B (2020). 'An Approach for Bibliographic Citation Sentiment Analysis Using Deep Learning'. 1 Jan. 2020: 353–362.

[17] Hossain E., Sharif O., Hoque M.M., Sarker I.H. (2021) SentiLSTM: A Deep Learning Approach for Sentiment Analysis of Restaurant Reviews. In: Abraham A., Hanne T., Castillo O., Gandhi N., Nogueira Rios T., Hong TP. (eds) Hybrid Intelligent Systems. HIS 2020. Advances in Intelligent Systems and Computing, vol. 1375. Springer, Cham. https://doi.org/10.1007/978-3-030-73050-5_19

[18] Rhanoui Maryem, Mikram Mounia, Yousfi Siham, Barzali Soukaina (2019). "A CNN-BiLSTM Model for Document-Level Sentiment Analysis" Mach. Learn. Knowl. Extr. 1, no. 3: 832–847. https://doi.org/10.3390/make1030048

[19] Devlin, J., Chang, M., Lee, K., Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. ArXiv, abs/1810.04805.

Biographies

Mifrah Sara received the bachelor's degree in Mathematical and Computer Science from Hassan II University of Casablanca FSBM in 2014, the master's degree in information science and engineering from the same University in 2016, and she is a PhD Student in computer science. Her research areas include Natural Language Processing, deep learning, and Bibliometric analysis. she has been serving as a reviewer for some journals.

El Habib Benlahmar holds a PhD in computer science from the National School of Computer Science and Systems Analysis in 2007. He is currently a professor of higher education at the Faculty of Sciences Ben M'Sik, Laboratory of Computer Science and Modeling, University Hassan II, Casablanca, Morocco. He has published several papers in various international journals and national and international conferences. His research interests include: Machine Learning, E-learning, Cloud Computing, Data Science, Ontology, Deep Learning, Internet of Things, Semantic Web, Bibliometric Analysis, Mathematics, Semantic Web Technologies, Mobile Applications, Educational Technology, Human-Computer Interaction.