Tarea 07 - Natural Language Processing
Giovanni Gamaliel López Padilla

1 Introduction

The sentiment analysis or mining opinion is the computational study about opinions, comments, emotions, feelings and actitude El análisis de sentimiento o opinion mining es el estudio computacional de las opiniones, comentarios, emociones, sentimientos y attitudes about entities as services, organizations, people, problems, events and topics.[1] The growth of the study of the sentiments in documents matchs with the social networks boom. The blogs, forums and twitter are include in this socialm networks. The sentiment analysis is one of the most active fields about Natural Language Processing (NLP). This tecnic has been used in other knowledge areas as managment sciende and social science as marketing, finance, politic science, communication and history.[2]

There are many methods that perform the sentiment analysis. This methods involves supervised learning and non supervised learning. The non supervised learning methods take advantage of lexicons, gramatical and syntactic patterns. The support vector machines (SVM), maximum entropy and naive bayes are examples for non supervised methods.[1,3,4]

At the beginning of the 2010s, Deep learning surges as an important method for the supervised learning.[5] This method produce results commensurable with the state-of-art on several applications in computer vision and speech recognition.

1.1 Sentiment analysis tasks

The sentiment analysis are divided in some levels.[6]

1.1.1 Document-level sentiment classification

In this level, the document is trated as primary information. The document is classificated with negative or positive label. Yang[7] proposed a hierarchical attention network model that focus in the information to build a document representation. The biggest challenge in document-level sentiment classification is create relationships with the words on a extense corpus. This problem can be trated with a SR-LSTM.[8] SR-LSTM model is composed with a LSTM that learn each word vector (sentence vectors) and a second layer enconde the link between words.

1.1.2 Sentence-level sentiment classification

The setence-level is used to classify if one word is objective or subjective. One word is classificated with subjective label if this word don't contain any sentiment. Zhao[9] proposed a framework called Weakly-supervised Deep Embedding (WDE). This framework was trained with review ratings and it's purpose to sentiment classifier using a Convolutional Neural Network (CNN). The architecture of this framework is two networks, WDE-CNN y (WDE-LSTM) to extract the review's vector representation. This model was evaluate with Amazon dataset from three domains (digital cameras, cell phones and laptops). The accuracy obtained on WDE-CNN model was 87.7%, and on WDE-LSTM model was 87.9%, which shows that deep learning models gives highest accuracy as compared to baseline models.

1.1.3 Aspect-level sentiment classification

Aspect level sentiment analysis is commonly called feature-based sentiment analysis or entity-based sentiment analysis. This sentiment analysis task includes the identification of features or aspects in a sentence (which is a user-generated review of an entity) and categorizing the features as positive or negative. The sentiment-target pairs are first identified, then they are classified into different polarities, and finally, sentiment values for every aspect are clubbed. Recently, attention-based LSTM mechanisms are being used for aspect-based sentiment analysis. Ma et al.[10] proposed a two-step attention architecture, which attends words of the target expression along with the whole sentence. The author also applied extended LSTM, which can utilize external knowledge for developing a common-sense system for target aspect-based sentiment analysis. The initial systems were not able to model different aspects in a sentence and do not explore the explicit position context of words. Hence, Ma et al.[11] developed a two-stage approach that can handle the above problems. In Stage-1, position attention model is introduced for modelling the aspects and its neighboring context words. In Stage-2 multiple aspect terms within a sentence are modelled simultaneously. The most recent approach is proposed by Yang et al.,[12] which replaces the conventional attention models with coattention mechanism by introducing a Coattention-LSTM network that can model the context-level and target-level attention alternatively by learning the non-linear representations of the target and context simultaneously. Thus, the proposed model can extract more effective sentiment features for aspect-based sentiment analysis.

1.1.4 Multi-domain sentiment classification

The word domain is referred as a set of documents that are related to a specific topic. Multi-domain sentiment classification focuses on transferring information from one domain to the next domain. The models are first trained in source domain. The knowledge is then transferred and explored in another domain. Yuan et al.[13] proposed a Domain Attention Model (DAM) for modeling the feature-level tasks using attention mechanism for multi-domain sentiment classification. DAM is composed of two modules: domain module and sentiment module. The domain mod- ule predicts the domain in which text belongs using bi-LSTM, and sentiment module selects the important features related to the domain using another bi-LSTM with attention mechanism. The vector thus obtained from the sentiment module is fed into a softmax classifier to predict the polarity of the texts. The author used Amazon multi-domain dataset containing reviews from four domains, and Sanders Twitter Sentiment dataset containing tweets about four different IT companies. The proposed model was compared with traditional machine learning approaches, and results show that the model performed well for multi-domain sentiment classification.

1.1.5 Multimodal sentiment classification

Different people express their sentiments or opinions in different ways. Earlier, the text was considered as the primary medium to express an opinion. This is known as a unimodal approach. With the advancement of technology and science, people are now shifting towards visual and audio modalities to express their sentiments. Combining or fusing more than one modalities for detecting the opinion is known as multimodal sentiment analysis. Hence, researchers are now focusing on this direction for improving the sentiment classification process. Poria et al.[14] proposed a novel methodology for merging the affective information extracted from audio, visual, and textual modalities. They discussed how different modalities were combined together to improve the overall sentiment analysis process. The experimental results showed that bimodal and trimodal models have shown better accuracy as compared to unimodal models, which shows the importance of using features from all the modality for enhancing the performance of sentiment analysis models.

1.1.6 Taxonomy of sentiment analysis

Research in the field of sentiment analysis is taking place for several years. Initially, handcrafted features were used for various classification tasks. On the other hand, machine-learned features can be categorized into traditional machine learning-based approaches and deep learning-based approaches. Machine learning-based methods include Support Vector Machine (SVM), Naïve Bayes (NB), Maximum Entropy (ME), Decision tree learning, and Random Forests. They are further categorized into supervised and unsupervised learning methods.

2 Dataset and strategies

2.1 Transformers

In the works of NLP, the use of pre-entrained language models hace become a useful block to get a better result on every task. One of the most competitive neural sequence transduction models have an encoder-decoder structure.[15,16] Here, the encoder maps an input sequence of symbol representations $(x_1, \dots, x_n)$ to a sequence of continuous representations $z = (z_1, \dots, z_n)$. Given z, the decoder then generates an output sequence $(y_1, \dots, y_m)$ of symbols one element at a time. At each step the model is auto-regressive,[17] consuming the previously generated symbols as additional input when generating the next. The Transformer follows this overall architecture using stacked self-attention and point-wise, fully connected layers for both the encoder and decoder (figure 1). The encoder is composed of a stack of N = 6 identical layers. Each layer has two sub-layers. The first is a multi-head self-attention mechanism, and the second is a simple, position wise fully connected feed-forward network.

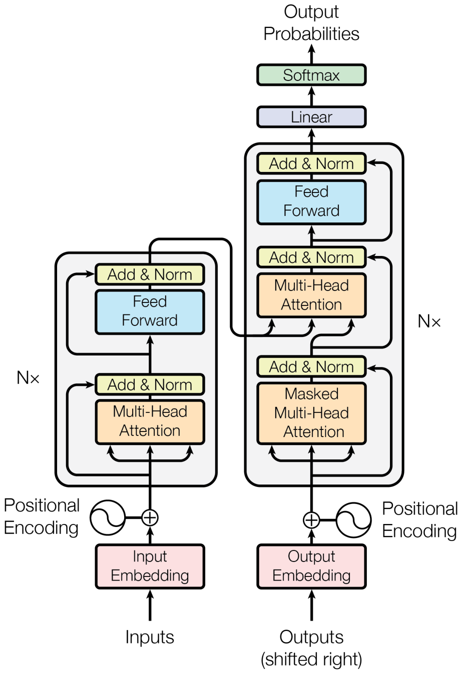

Figure 1: Transformer model representation.[18]

2.2 BERT

BERT model is an acronym for Bidirectional Encoder Representations for Transformers. BERT alleviates the previously mentioned unidi rectionality constraint by using a 'masked lan guage model' (MLM) pretraining objective, in spired by the Cloze task.[19] The masked language model randomly masks some of the tokens from the input, and the objective is to predict the original vocabulary id of the masked word based only on its context. Pre-trained word embeddings are an integral part of modern NLP systems, offering significant improvements over embeddings learned from scratch.[20] To pre-train word embedding vectors, left-to-right language modeling objectives have been used,[21] as well as objectives to discriminate correct from incorrect words in left and right context.[22] As with the feature-based approaches, the first works in this direction only pre-trained word embedding parameters from unlabeled text.[23] More recently, sentence or document encoders which produce contextual token representations have been pre-trained from unlabeled text and fine-tuned for a supervised downstream task. The advantage of these approaches is that few parameters need to be learned from scratch.

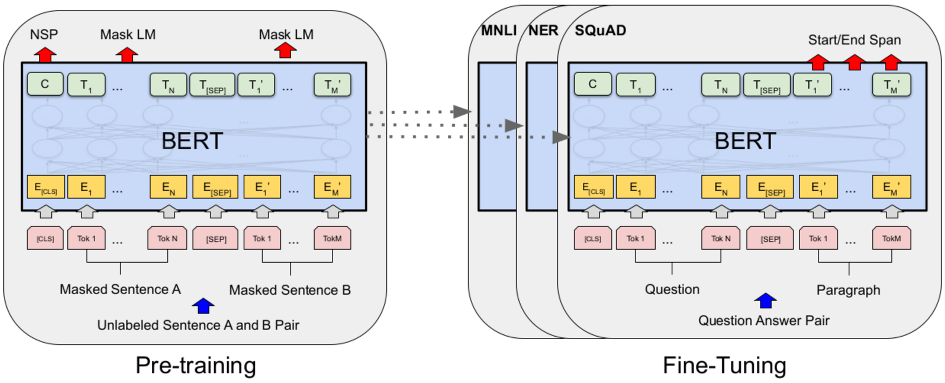

Figure 2: Overall pretraining and fine-tuning procedures for BERT. Apart from output layers, the same architectures are used in both pretraining and fine-tuning. The same pre-trained model parameters are used to initialize models for different down-stream tasks.[24]

There are two steps in our framework: pretraining and fine-tuning. During pretraining, the model is trained on unlabeled data over different pretraining tasks. For fine-tuning, the BERT model is first initialized with the pre-trained parameters, and all of the parameters are fine-tuned using labeled data from the downstream tasks. Each downstream task has separate fine-tuned models, even though they are initialized with the same pre-trained parameters. The question-answering example in Figure 2 will serve as a running example for this section. To make BERT handle a variety of down-stream tasks, our input representation is able to unambiguously represent both a single sentence and a pair of sentences in one token sequence. Throughout this work, a sentence can be an arbitrary span of contiguous text, rather than an actual linguistic sentence.

2.3 RoBERTa

The BERT model can be optimized with some modifications on the pretraining procedure. Liu[25] join this configurations in one model named Robustly optimized BERT approach (RoBERTa). Especifically, RoBERTa is trained with dynamic masking, FULL-SENTENCES without NS loss, large mini-batches. One of the most important modifications is the number of training passes and the size of the bacth. This is because Large batch training can improve training efficiency even without large scale parallel hardware through gradient accumulation, whereby gradients from multiple mini-batches are accumulated locally before each optimization step.[26] In all the test that Liu[25] did in his paper demostrate that RoBERTa have a perfomance by training the model with bigger batches over more data, removing the next sentence prediction, training on longer sequences and dynamically changing the masking pattern applied to the traning data.

2.4 RoBERTuito

The RoBERTuito model has a RoBERTa base architecture. This model have 2 self-attention layers, 12 attention heads, and hidden size equal to 768, in the same fashion as BERTweet.[27]

RoBERTuito use a masked language objective disregarding the next-sentence prediction task used in BERT or other tweet-order tasks such as those used in Gonzalez et al.[28]

2.5 MEX-A3T

The MEX-A3T is an evaluation forum for IberLEF intended for the research in NLP and considering a variety of Mexican Spanish cultural traits. In this vein, the 2018 edition was the first to consider the aggressiveness identification for Mexican Spanish tweets.[29] This dataset have two columns with 7332 rows (5278 for train, 587 for validation and 1467 for test). The categories are offensive (1) and no-offensive (0). The distribution of this categorias in the data are show in figure 3.

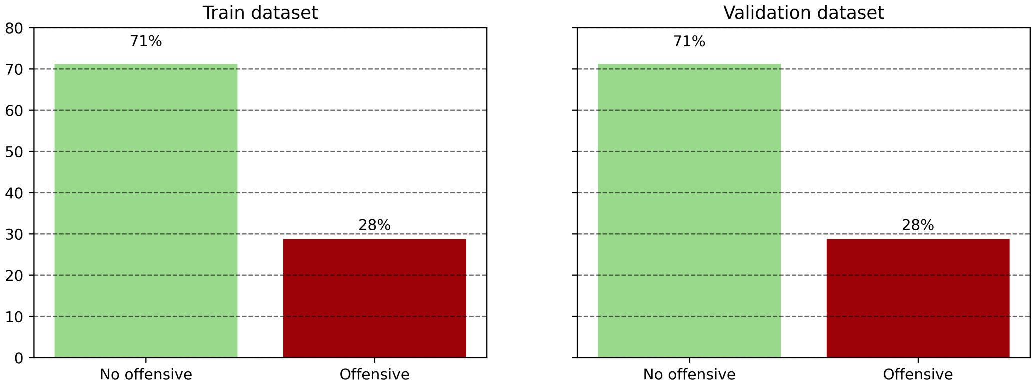

Figure 3: Distribution of the categories (offensive and no offensive) for the train and validation dataset from MEX-A3T.

In the Iberian Languages Evaluation Forum 2020 (IberLEF 2020) 21 teams participated. The evaluation consisted in two task, fake news track and aggressiveness identification. The results of this evaluation is in table 1 and 2.

Team Name	Fake	Truth	F1 macro	Precision	Recall	Accuracy
Idiap-UAM-2	0.8444	0.86088	0.85660	0.8615	0.8557	0.85760
Idiap-UAM-1	0.8406	0.85990	0.85020	0.8521	0.8496	0.85080
Ares	0.8188	0.81510	0.81690	0.8191	0.8185	0.81690
CIMAT-1	0.7943	0.81170	0.80300	0.8032	0.8029	0.80340
Baseline (BoW-RF)	0.7850	0.78790	0.78640	0.7870	0.7873	0.78640
Intensos-2	0.7703	0.78830	0.77930	0.7794	0.7792	0.77970
Intensos-1	0.7597	0.73760	0.74870	0.7555	0.7518	0.74920
INGEOTEC	0.7596	0.77230	0.76590	0.7659	0.7662	0.76061
ITCG-SD	0.7464	0.77710	0.76017	0.7632	0.7614	0.76270

Table 1: Results for Fake News track in the IberLEF 2020.

Team Name	F1 offensive	F1 non-offensive	F1 macro	Precision	Recall	Accuracy
CIMAT-1	0.7998	0.9195	0.8596	0.8605	0.8588	0.8851
CIMAT-2	0.7971	0.9205	0.8588	0.8641	0.8540	0.8858
UPB-2	0.7969	0.9107	0.8538	0.8440	0.8668	0.8759
UACh-2	0.7720	0.9042	0.8381	0.8332	0.8437	0.8651
INGEOTEC	0.7468	0.8933	0.8200	0.8150	0.8258	0.8498
Idiap-UAM-1	0.7255	0.8886	0.8071	0.8067	0.8075	0.8416
Baseline (Bi-GRU)	0.7124	0.8841	0.7983	0.7988	0.7978	0.8348
Idiap-UAM-2	0.7066	0.8953	0.8010	0.8234	0.7860	0.8457
UACh-1	0.7062	0.8861	0.7961	0.8021	0.7909	0.8358
DeepMath-1	0.7001	0.8544	0.7773	0.7662	0.8120	0.8040
DeepMath-2	0.6957	0.8537	0.7747	0.7639	0.7971	0.8024
Baseline (BoW-SVM)	0.6760	0.8780	0.7770	-	-	0.8228
UMUTeam-2	0.6727	0.8706	0.7716	0.7744	0.7691	0.8145
Intensos-1	0.6619	0.8752	0.7686	0.7820	0.7588	0.8177
UMUTeam-3	0.6516	0.8771	0.7644	0.7868	0.7503	0.8183
Ugalileo-2	0.6388	0.8208	0.7298	0.7213	0.7531	0.7604
Ugalileo-1	0.6387	0.8430	0.7408	0.7350	0.7486	0.7811
ITCG-SD	0.6080	0.8820	0.7450	0.8133	0.7203	0.8186
UMUTeam-1	0.5892	0.8430	0.7161	0.7223	0.7112	0.7728
UPB-1	0.3437	0.8463	0.5950	0.7333	0.5947	0.7509
Intensos-2	0.2515	0.7664	0.5090	0.5189	0.5141	0.6440

Table 2: Results for aggressiveness identification in the IberLEF 2020.

2.6 Implementation

The implementation of this work is based on RoBERTuito uncased model. The loss function used was Croos Entropy from pytorch library. This function recives a tensor with weights from the data. This procedure benefits the unbalansed clategories data (figure 3). The optimization method was AdaW from pytorch library. The hyperparameters chosed for this report are in the table 3.

Hyperparameter	Value
Batch size	8
Epochs	3
Learning rate	1x10^{-5}
Max tokens	130

Table 3: Hyperparameters used in the implementation.

3 Results

The training and validation history of loss function and accuracy for RoBERTuito is in figure 4. This model with the hyperparameters from table 3 obtain a 0.89143 on F1-Score Macro on the test dataset. This value was calculatd with the Kaggle Forum. With a bigger number on epochs, the F1-Score is reduced to 0.82 or 0.80. The learing rate (lr) can be 2x10^{-5}, this value can produce same F1-Scores, but if the lr increase or decrease more the F1-Score will be lower. The model was based in this blog How to Train BERT.

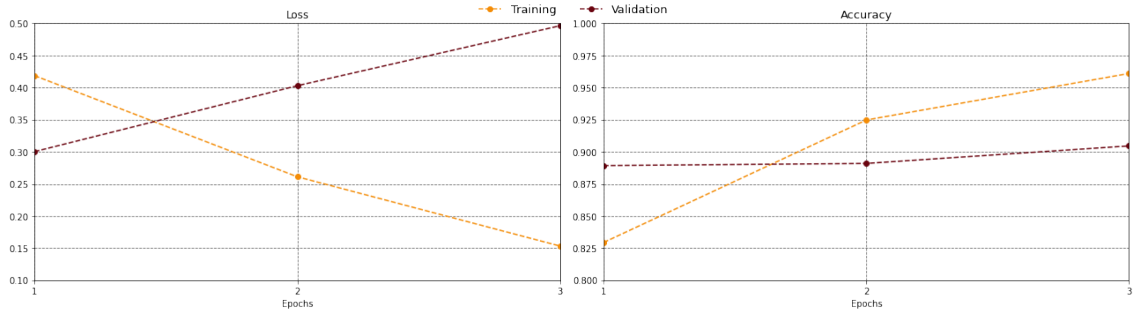

Figure 4: Training and validation history of loss function and accuracy.

4 Conclusions

The RoBERTuito model have a better result for the sentiment analysis. The RNN, CNN, Bert, RoBERTa models have a lower results than nlp_cimat, but only BERT and RoBERTa make a prediction better than FerdotSV. The HeatEval dataset was taken to create apply data augmentation, but i think there is an character that raises an error in the GPU from Google Colab. RoBERTuito have a better result than the others models because the pretrained base was excecute wtih a spanish tweet database.

5 References

[1] Liu B. Sentiment Analysis: Mining Opinions, Sentiments, and Emotions. Cambridge University Press; 2015.

[2] Chalothom T, Ellman J. Simple Approaches of Sentiment Analysis via Ensemble Learning. In: Lecture Notes in Electrical Engineering. Springer Berlin Heidelberg; 2015. p. 631–639. Available from: https://doi.org/10.1007%2F978-3-662-46578-3_74.

[3] Liu B. Sentiment Analysis and Opinion Mining. Synthesis Lectures on Human Language Technologies. 2012 may;5(1):1–167. Available from: https://doi.org/10.2200%2Fs00416ed1v01y201204hlt016.

[4] Lee H, Grosse R, Ranganath R, Ng AY. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In: Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09. ACM Press; 2009. Available from: https://doi.org/10.1145%2F1553374.1553453.

[5] Goodfellow I, Bengio Y, Courville A. Deep Learning. MIT Press; 2016. Available from: http://www.deeplearningbook.org.

[6] Thomas B. What Consumers Think about brands on social media, and what bunesses need to do about it. Report, Keep Social Honest. 2013;Available from: https://www.cim.co.uk/media/1733/ksh_report_.pdf.

[7] Yang Z, Yang D, Dyer C, He X, Smola A, Hovy E. Hierarchical Attention Networks for Document Classification. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego, California: Association for Computational Linguistics; 2016. p. 1480–1489. Available from: https://aclanthology.org/N16-1174.

[8] Rao G, Huang W, Feng Z, Cong Q. LSTM with sentence representations for document-level sentiment classification. Neurocomputing. 2018;308:49–57. Available from: https://www.sciencedirect.com/science/article/pii/S092523121830479X.

[9] Zhao W, Guan Z, Chen L, He X, Cai D, Wang B, et al. Weakly-Supervised Deep Embedding for Product Review Sentiment Analysis. IEEE Transactions on Knowledge and Data Engineering. 2018;30(1):185–197.

[10] Ma Y, Peng H, Khan T, Cambria E, Hussain A. Sentic LSTM: a Hybrid Network for Targeted Aspect-Based Sentiment Analysis. Cogn Comput. 2018 mar;10(4):639–650. Available from: https://doi.org/10.1007%2Fs12559-018-9549-x.

[11] Ma X, Zeng J, Peng L, Fortino G, Zhang Y. Modeling multi-aspects within one opinionated sentence simultaneously for aspect-level sentiment analysis. Future Generation Computer Systems. 2019;93:304–311. Available from: https://www.sciencedirect.com/science/article/pii/S0167739X18319125.

[12] Yang C, Zhang H, Jiang B, Li K. Aspect-based sentiment analysis with alternating coattention networks. Information Processing & Management. 2019;56(3):463–478. Available from: https://www.sciencedirect.com/science/article/pii/S0306457318306344.

[13] Yuan Z, Wu S, Wu F, Liu J, Huang Y. Domain attention model for multi-domain sentiment classification. Knowledge-Based Systems. 2018;155:1–10. Available from: https://www.sciencedirect.com/science/article/pii/S0950705118302144.

[14] Poria S, Cambria E, Howard N, Huang GB, Hussain A. Fusing audio, visual and textual clues for sentiment analysis from multimodal content. Neurocomputing. 2016;174:50–59. Available from: https://www.sciencedirect.com/science/article/pii/S0925231215011297.

[15] Bahdanau D, Cho K, Bengio Y. Neural Machine Translation by Jointly Learning to Align and Translate. arXiv; 2014. Available from: https://arxiv.org/abs/1409.0473.

[16] Cho K, van Merrienboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, et al.. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. arXiv; 2014. Available from: https://arxiv.org/abs/1406.1078.

[17] Graves A. Generating Sequences With Recurrent Neural Networks. arXiv; 2013. Available from: https://arxiv.org/abs/1308.0850.

[18] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is All you Need. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, et al., editors. Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc.; 2017. Available from: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[19] Taylor WL. “Cloze Procedure”: A New Tool for Measuring Readability. Journalism Quarterly. 1953 sep;30(4):415–433. Available from: https://doi.org/10.1177%2F107769905303000401.

[20] Turian J, Ratinov L, Bengio Y. Word Representations: A Simple and General Method for Semi-Supervised Learning. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. ACL ’10. USA: Association for Computational Linguistics; 2010. p. 384–394.

[21] Mnih A, Hinton GE. A Scalable Hierarchical Distributed Language Model. In: Koller D, Schuurmans D, Bengio Y, Bottou L, editors. Advances in Neural Information Processing Systems. vol. 21. Curran Associates, Inc.; 2008. Available from: https://proceedings.neurips.cc/paper/2008/file/1e056d2b0ebd5c878c550da6ac5d3724-Paper.pdf.

[22] Le QV, Mikolov T. Distributed Representations of Sentences and Documents. arXiv; 2014. Available from: https://arxiv.org/abs/1405.4053.

[23] Collobert R, Weston J. A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning. ICML '08. New York, NY, USA: Association for Computing Machinery; 2008. p. 160–167. Available from: https://doi.org/10.1145/1390156.1390177.

[24] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv; 2018. Available from: https://arxiv.org/abs/1810.04805.

[25] Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al.. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv; 2019. Available from: https://arxiv.org/abs/1907.11692.

[26] Ott M, Edunov S, Baevski A, Fan A, Gross S, Ng N, et al.. fairseq: A Fast, Extensible Toolkit for Sequence Modeling. arXiv; 2019. Available from: https://arxiv.org/abs/1904.01038.

[27] Nguyen DQ, Vu T, Tuan Nguyen A. BERTweet: A pre-trained language model for English Tweets. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. Online: Association for Computational Linguistics; 2020. p. 9–14. Available from: https://aclanthology.org/2020.emnlp-demos.2.

[28] Ángel González J, Hurtado LF, Pla F. TWilBert: Pre-trained deep bidirectional transformers for Spanish Twitter. Neurocomputing. 2021;426:58–69. Available from: https://www.sciencedirect.com/science/article/pii/S0925231220316180.

[29] Carmona MA, Guzmán-Falcón E, Montes M, Escalante HJ, Villaseñor-Pineda L, Reyes-Meza V, et al. Overview of MEX-A3T at IberEval 2018: Authorship and aggressiveness analysis in Mexican Spanish tweets; 2018. .

[30] Zhang L, Wang S, Liu B. Deep learning for sentiment analysis: A survey. WIREs Data Mining Knowl Discov Data Mining and Knowledge Discovery. 2018 mar;8(4). Available from: https://doi.org/10.1002%2Fwidm.1253.

[31] Zhao J, Liu K, Xu L. Sentiment Analysis: Mining Opinions, Sentiments, and Emotions. Computational Linguistics. 2016 09;42(3):595–598. Available from: https://doi.org/10.1162/COLI_r_00259.

[32] Yessenalina A, Yue Y, Cardie C. Multi-Level Structured Models for Document-Level Sentiment Classification. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. Cambridge, MA: Association for Computational Linguistics; 2010. p. 1046–1056. Available from: https://aclanthology.org/D10-1102.

[33] Farra N, Challita E, Assi RA, Hajj H. Sentence-Level and Document-Level Sentiment Mining for Arabic Texts. In: 2010 IEEE International Conference on Data Mining Workshops; 2010. p. 1114–1119.

[34] Engonopoulos N, Lazaridou A, Paliouras G, Chandrinos K. ELS. In: Proceedings of the International Conference on Web Intelligence, Mining and Semantics - WIMS '11. ACM Press; 2011. Available from: https://doi.org/10.1145%2F1988688.1988703.

[35] Zhou H, Song F. Aspect-Level Sentiment Analysis Based on a Generalized Probabilistic Topic and Syntax Model. In: FLAIRS Conference; 2015. Available from: https://www.semanticscholar.org/paper/Aspect-Level-Sentiment-Analysis-Based-on-a-Topic-Zhou-Song/262d4a59b557c109255ae89cab3c927cde44773d.

[36] Pérez JM, Furman DA, Alemany LA, Luque F. RoBERTuito: a pre-trained language model for social media text in Spanish. arXiv; 2021. Available from: https://arxiv.org/abs/2111.09453.