# Sentiment Analysis Based on Bert for Amazon Reviewer

Mohamed Mustafa[a] and Asmaa Elsaid[b]

[a] Higher Institute of Computers and Information Technology, Computer Dept., El. Shorouk Academy, Cairo, Egypt, Email: Dr.Mohamed.Mustafa@sha.edu.eg

[b] Higher Institute of Computers and Information Technology, Computer Dept., El. Shorouk Academy, Cairo, Egypt, Email: Asmaa.alsaeed@sha.edu.eg

## Abstract

Sentiment analysis determines if a text includes subjective information and what that information represents, i.e., whether the text's attitude is positive, negative, or neutral. Understanding user-generated content sentiments automatically help commercial and political interests. Classify the polarity of words, phrases, or entire documents. The demand for sentiment analysis is raised due to the requirement of analyzing and structuring hidden information, extracted from Amazon reviews in form of unstructured data. The sentiment analysis is being implemented through deep learning, machine learning, and lexicon techniques. In the research, multiple machine learning algorithms are evaluated, trained, and tested using Amazon product reviews randomly picked from a 4 million-review Kaggle dataset. The performance of nine different algorithms was compared: KNN, Decision Tree, Naive Bayes, Random Forest, Logistic Regression, SVM, Bidirectional LSTM, GRU, and Bert to reach the highest performance (accuracy). The Bert resulted in the highest performance with an Accuracy of 0.94. Thereafter, to evaluate the Bert model, it was applied to 502,103 reviews, split into a 90% train set to train the model and a 10% test set to evaluate the Bert mode. It has been proven that Bert networks are very suitable for the classification of sentiment in product reviews.

**Keywords:** NLP, Sentiment analysis, deep learning, Bert, Amazon

## *1. Introduction*

In today's world, the Amazon.com portal is the best or most popular way to show customers' feelings or sentiments by writing reviews. Customers buy products and write product reviews. Every person in this world shows their attitude or feelings by writing comments, reviews, etc. Here comes the role of sentiment analysis: to study the feelings or sentiments of the customers. Sentiment analysis, also known as opinion mining, is the field of data mining that studies a person's attitude, behavior, and feelings toward objects, organizations, events, products, etc. Sentiment analysis has four types: polarity (fine-grained) sentiment analysis, emotion detection, aspect sentiment analysis, and multilingual sentiment analysis.

The internet is regarded as one of the most important sources of consumer opinion, enabling the release of several websites. Customers can offer their ratings and thoughts about many things on these websites, including films, eateries, hotels, gadgets, and books. Amazon, which offers millions of users evaluations of various product categories, is an example. of the increasing availability and popularity of opinion-rich resources. And it's no different here in Egypt, with more than

8.9 million visits "only last month," which makes it a massive market that has plenty of opportunities [23]. But our question is how Many of those customers are really satisfied with the services. that they are getting? We all know the Amazon book got a rating. system of its own for each product, but it is not globally available. their entire website, which we are aiming to do.

Sentiment analysis is the contextual mining of words, which indicates the social media analysis of feedback or reviews. regarding the brands or products, which helps marketers determine whether

their product is going to attract demand. in the market or not. Is a data mining technique that uses NLP, computational linguistics, and text analysis to identify and extract the content of interest from a text's body, which can help measure customer satisfaction on Amazon.

Sentiment analysis has three approaches: rule-based, automatic, and hybrid, which combine rule-based and automatic [1] [2]. These automated systems do sentiment analysis based on a predetermined set of rules that were developed by humans using rule-based techniques. To learn from data automatically, automated methods use machine learning techniques.

The contribution of this paper is a customer satisfaction model and a comparison of state-of-the-art methods. The main objectives of this paper are:

Measuring customer satisfaction at Amazon, e.g. To achieve this, this study aims to scrape customer reviews from Amazon, e.g., and apply it to preprocessing like dealing with missing data and eliminating linkages, tags, numerical values, stop words, and more cleaning techniques Then applying Bert to predict sentiment on the Amazon dataset (a huge dataset of 4 million reviews). The experimental results were compared with other deep learning approaches such as bi-directional long-term memory (Bi-LSTM), gated recurring units (GRU), support vector machines (SVM), logistic regression, random forest, Naive Bayes, Decision Tree, and KNN, and visualizing data and drawing conclusions. The proposed model using Bert was evaluated by F-measure with an accuracy of 0.94.

This paper is divided into Section 2 will present the background., Section 3 presents related work, which includes earlier work that addressed the same issue as ours. Section 4 which presents the methodology, consisting of dataset specifications, data collection, preprocessing, and Sentiment and Evaluation Also, the results are discussed in Section 5, and finally, Section 6 is for the conclusion.

## *2. Background*

Sentiment analysis is the process of analyzing products. reviews on the internet to determine the overall opinions and expressions about a product, so the text of these opinions and expressions could be classified as positive, negative, or neutral. It specifically focuses on evaluating the opinions and expressions on a topic of interest using machine learning. Techniques. The machine learning approach, which is an automatic approach to sentiment analysis, is widely used for Sentiment classification and is different from the linguistic methods approach, which is a rule-based approach to sentiment analysis.

However, just classifying concepts as positive or negative is insufficient. There are several challenges to overcome. Positive and negative polarity cannot always be used to classify words and sentences. For example, the word "amazing" used to have a positive connotation, but when coupled with a negative word like "not," the meaning might radically shift. Emotion classification has been tried in a variety of contexts, including product reviews, movie reviews, and hotel reviews. To identify feelings, machine learning methods are frequently utilized.

However, these sentiment analysis approaches don't perform well. with the same efficiency of sentiment classification in topic categorization. Since the nature of the opinionated text requires more understanding of the text, machine learning classifiers such as Naive Bayes, maximum entropy, and support. Vectors are used for sentiment classification to achieve high accuracy of categorization.

The feedback or reviews, which are user-generated content, are a rich source for marketing specialists who are concerned with public moods and the personal attitudes of the customers toward what is offered by the marketer through brands and products. Due to the diversity and size of social media data, sentiment analysis is applied instead of collecting data. manually through individuals or companies, as it is an automated and real-time opinion extraction and mining system. Customers can submit evaluations on a wide variety of products on e-commerce websites, which is why they are growing increasingly

popular [11]. Every day, millions of reviews are written by customers, making it tough for manufacturers to keep track of their thoughts on the product. As a result, it is critical to disperse massive and complicated data in order to extract usable information from them.

Classification algorithms are the best technique to deal with such problems. Classification is the process of dividing data into groups or classes based on common qualities. The capacity to automate the classification process when working with enormous datasets is a major concern for businesses [12].

Sentiment analysis, also known as opinion mining, is a type of natural language processing (NLP) task that entails subjective data extraction from text sources and abstract data. The goal of emotion classification is to look at user feedback and identify it as good or negative. This avoids the requirement for the system to completely comprehend each sentence's semantics.

Bidirectional Representation for Transformers, or BERT, is a pre-trained language model that is designed to consider the context of a word from both the left and right sides simultaneously. It improves results at several NLP tasks, including sentiment analysis and question-and-answer systems. As a pre-trained language model, BERT provides context to words for representing them from unannotated training data. So, it could extract more context features from a sequence compared to viewing the left and right sides simultaneously. BERT is adaptable to perform different NLP tasks with state-of-the-art accuracy, similar to the transfer learning method in computer vision, which allows for building accurate models in a time-saving way.

## *3. Related works*

Sentiment Analysis BERT-based models are effectively used in many natural language processing tasks such as the sentiment analysis task. A common procedure is to start from an off-the-shelf pre-trained model and then fine-tune it on a specific task. However, in many tasks, including sentiment analysis, the fine-tuning needs labeled data, which could be lacking for specific domains such as health or finance. Nevertheless, little research has been conducted on the performance variability of BERT-based models for sentiment analysis over multi-domain and multi-source corpora to investigate the universal applicability of these models. Moreover, evaluations of BERT-based models for the Italian language report, to the best of our knowledge, only the performance on positive and negative classes, although the class neutral is needed since it identifies the absence of a predominant and clear attitude. A multi-domain corpus for sentiment analysis for the English language is presented by [13]. This corpus is a collection of tweets gathered using a set of keywords, which were selected to identify various socially relevant domains. However, the evaluated models are non-BERT-based and the evaluation does not report the performances for each domain. In [15], the authors conducted research to perform the classification of customer reviews, followed by finding the sentiment of the reviews, to provide visualization and summarization for the results. The classification of reviews was done along with sentimental analysis, which provided accurate reviews to the user. In [16], the authors conducted research to examine the effectiveness of different machine-learning techniques for the classification of online reviews using supervised learning methods, and also the extraction of product feature perception for deducing adjective polarity when the polarity is unknown. Sentiment analysis was used to gather a lot of information, and this information was where these training data were previously gathered. The results from the perception of product features subtask did not have sufficient test data. where the subtask was more complicated than the document-level sentiment analysis, but more care should be taken to give verifiable results. The results from the polarity deduction subtask were something of an afterthought compared with the other subtasks conducted in the research. In [17], research to polarize the feedback of customers over different products, which was done through the supervised learning method, is conducted on a large-scale Amazon dataset to polarize it and get satisfactory accuracy. The sentiment analysis resulted in an accuracy rate of

over 90%. Different simulations were applied using cross-validation, training-test ratios, and different feature extraction processes for comparing varying amounts of data. In [14], the authors conducted research to implement and test Amazon customer reviews where aspect terms are identified first for each review. The system performs preprocessing operations to extract meaningful information, so meaningful information could be extracted and classified as either positive or negative. Identification of words changing polarity took place in the presence of context, and its effect on the overall rating of the product, along with the aspect, has been analyzed in the work.

## *4. Methodology*

The proposed sentiment analysis is based on Bert on Amazon Reviewer. In this section, the stages of the proposed model, including data collection, data exploration, data preprocessing, feature extraction, training model, and evaluation metric, are discussed as shown in figure 1.

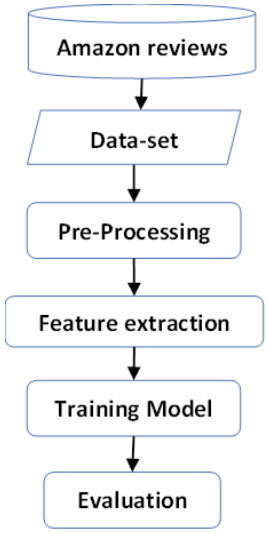


*Figure 1: The Architecture of the proposed system*

### *4.1 Data Collection*

Data is the most valuable item in the world now, and getting it is not as easy as possible. Amazon provides millions of data from their users' reviews of their products, so we relied on Amazon's review data set to train and evaluate our model. We collected a dataset from Kaggle [3]. This dataset includes 3.6 million training reviews and 0.4 million testing reviews from Amazon customers, as well as labels (output labels) for the reviews

### *4.2 Data Exploration*

Through exploring a dataset, we get a decent grasp of its structure and contents, making future navigation and usage more accessible. One's ability to analyze data depends on how well one is familiar with that data. For example, we read the dataset, took a sample review of nearly 250,000 reviews, and used a Count plot to determine whether the sentiments in the dataset are balanced or not as shown in figure 2.

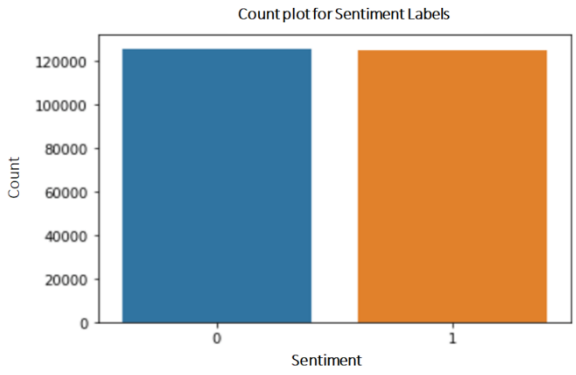


*Figure 2: count plot of sentiment Analysis Labels*

### *4.3 Data Preprocessing*

Preprocessing is preparing the dataset to have excellent quality. These are important tasks that must take place before using a dataset for model training. In order for the models to be professionally trained and to provide the expected results, the data used must be representative, clean, precise, complete, and well-labeled as shown in figure 3:

- Removal of HTML tags.
- Removal of NULL and repeated data from the dataset
- Filtering every symbol except for letters (a-z) and numbers (0–9).
- Filtering out every word with a length of 3 symbols or lower
- Removal of stop words
- Applied tokenized by splitting sentences into words
- Applied stemming
- Applied spell correction
- Applied lemmatization
- Detect and delete non-English reviews

### *4.4 Feature Extraction*

The fundamental issue when dealing with language processing is that machine learning algorithms cannot be applied directly to raw text. Therefore, we need feature extraction methods that can transform the text into a feature matrix (or vector).

There are two different kinds of feature extraction available here: basic feature extraction and high-level feature extraction. Common ways of extracting features, like TF-IDF and bag-of-words, are not very dense, but there are many ways to get a dense vector representation of the words.

Extraction of advanced features (word embeddings) Word embeddings are a representation of words as reduced-dimensional vectors of numbers.

Because word embedding vectors are meant to be representations of both words and their surroundings, it stands to reason that words that have similar meanings (synonyms) or close semantic links would also share similar embeddings. For example, "man" should be "King" while "woman" is "queen". Pre-trained incorporations were used since learning word embeddings requires time and computation.

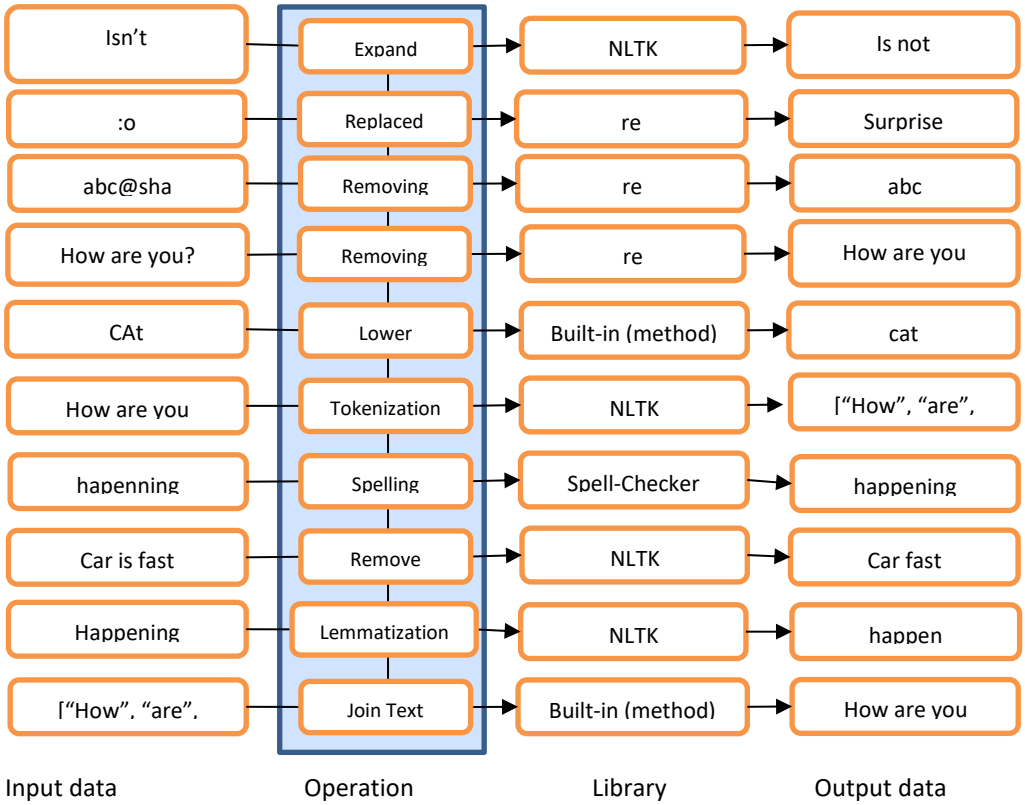


*Figure 3: Example of the basic operation of preprocessing and applying it to a built-in method in Python*

### *4.5 Training the model*

We applied many different deep learning models like KNN, Decision Tree, Naive Bayes, Random Forest, Logistic Regression, SVM, Bidirectional LSTM, GRU, and BERT to predict the sentiment. We have two output classes: positive sentiments like "happy," "in love," "smile," etc., and negative sentiments like "afraid," "not happy," and "angry. Our main interest was to use BERT's algorithm [4], and we made four attempts to get high accuracy. The attempts like change in hyperparameters like learning rate, and finally we used (1e-5) LR, increased the dataset, and finally we took a 502,103 sample of reviews. We activated all the non-trainable parameters, and we split the dataset into 90% to train the model and 10% to test the model. Machine learning models used a 10,000 sample of reviews and GRU and Bi-LSTM used 100,000. Bert Preprocessing: We have used preprocessing API from (TensorFlow Hub) to Clean and prepare the dataset to be usable in the model. Bert Encoder We have used Encoder API from (TensorFlow Hub) to transfer each word into a vector so the machine can understand. Splitting dataset Data is split into 3 sets: the Training set, Validating set, and the Test set.

## 5. EXPERIMENTAL RESULTS AND EVALUATION

### 5.1 Experimental setting

The model is trained using the COLLAB platform. Using Python 3 and a built-in library of NLP such as Pandas, it is used for data analysis and manipulation. Using NLTK, it is used It contains predefined libraries for preprocessing textual data. Vader, which is used to calculate sentiment scores, gives opinion scores in terms of positive, negative, and neutral. Numpy is used to deal with high-level mathematical functions to operate on arrays. While Seaborn and Matplotlib are used to visualize data, Lang-id is used to detect languages in Sklearn, used in machine learning.

### 5.2 Results and Discussions

In the first iteration, we used a BERT version called "BERT small" [5], which consists of four layers and 28 million parameters. And we provide the model with 50,000 reviews to train our model. In this model, we see that the number of trainable parameters is too small. So that was a problem. So, in this model, we get 50% accuracy. In the second iteration, we were able to activate those non-trainable parameters in the model by increasing the number of reviews to 250,000 reviews. We reached an accuracy of 84%. In the third iteration, we managed to push it further after reaching 84%. By increasing the dataset to 502,103 reviews, split into a 90% train set and a 10% test set, and using the same hyperparameters, we reached 92% accuracy. For the fourth iteration, 92% wasn't enough for us, so we decided to increase it, using the same model in the second iteration. But we trained it with 502,103 reviews, split it into a 90% train set and a 10% test set, and used the same hyperparameter in the fourth iteration. So, we ended up with an accuracy of 94% as shown in figure 4. Model Evaluation quantifies a system's predictions. The confusion matrix [6] helps solve classification issues and assess models. Figure 3 shows a confusion matrix used for binary and multiclass classification as shown in figure 5. After applying some machine learning and deep learning models, we can get the results of each model. The final result shows that Bert's model gives high accuracy with optimization time, as shown in Table 1.

| Model | Accuracy |
|---|---|
| KNN [7] | 66% |
| Decision Tree [7] | 69% |
| Naïve Bayes [7] | 79% |
| Random Forest [8] | 80% |
| Logistic Regression [9] | 81.9% |
| SVM [9] | 82.4% |
| Bidirectional Lstm [10] | 82.3% |
| GRU [10] | 86% |
| Bert | 94% |

*Table 1: The comparison of different deep learning models on the Amazon reviewer data set*

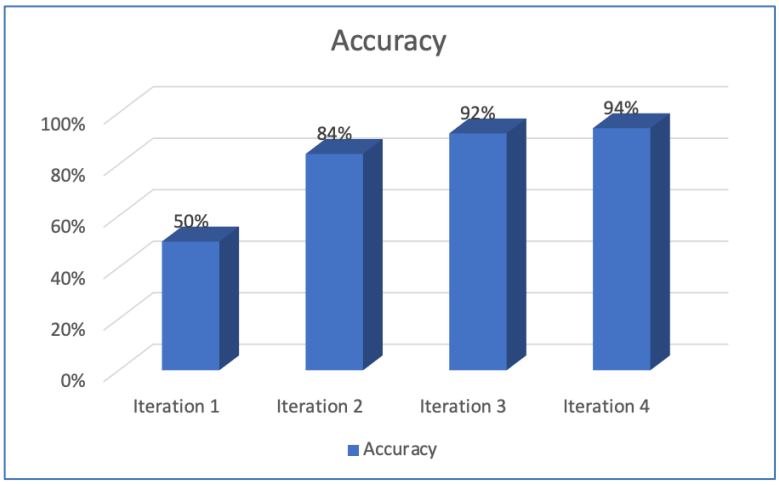


*Figure 4: the training accuracy*

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.93 | 0.94 | 0.94 | 25293 |
| 1 | 0.94 | 0.93 | 0.93 | 24917 |
| accuracy | | | 0.94 | 50210 |
| macro avg | 0.94 | 0.94 | 0.94 | 50210 |
| weighted avg | 0.94 | 0.94 | 0.94 | 50210 |

*Figure 5: Bert's Model accuracy*

## 6. Conclusion and future work

As shown in this paper, the BERT networks are the most suitable for binary sentiment analysis on Amazon.com product reviews. Based on the results of the evaluation datasets, we can conclude that BERT performs very well, with an accuracy of 0.94 for binary classification, and that does not depend strongly on the type of product the reviews come from. As can be seen clearly from the previous confusion matrices, the BERT network performs both accurate results for negative and positive classes. Since the training dataset is also balanced, getting balanced results from both classes shows the model's reliability. If we want to develop Web and mobile apps, we may use these models. Complex systems communicate to achieve a goal. We'll utilize Flask to build our model into a web API. Deep learning models that use Bart to improve results and Blue as an evaluation metric

## 7. References

[1] Birjali, Marouane, Mohammed Kasri, and Abderrahim Beni-Hssane. "A comprehensive survey on sentiment analysis: Approaches, challenges, and trends." Knowledge-Based Systems 226 (2021): 107134.

[2] Wankhade, Mayur, Annavarapu Chandra Sekhara Rao, and Chaitanya Kulkarni. "A survey on sentiment analysis methods, applications, and challenges." Artificial Intelligence Review (2022): 1-50.

[3] https://www.kaggle.com/datasets/kritanjalijain/amazon-reviews/code

[4] Alaparthi, Shivaji, and Manit Mishra. "BERT: A sentiment analysis odyssey." Journal of Marketing Analytics 9.2 (2021): 118-126.

[5] Bhargava, Prajjwal, Aleksandr Drozd, and Anna Rogers. "Generalization in NLI: Ways (not) to go beyond simple heuristics." arXiv preprint arXiv:2110.01518 (2021).

[6] Sajid, Muhammad, et al. "Impact of Land-use Change on Agricultural Production & Accuracy Assessment through Confusion Matrix." (2022).

[7] Zerrouki, Kadda, Reda Mohamed Hamou, and Abdellatif Rahmoun. "Sentiment Analysis of Tweets Using Naïve Bayes, KNN, and Decision Tree." Research Anthology on Implementing Sentiment Analysis Across Multiple Disciplines. IGI Global, 2022. 538-554.

[8] Li, Cai, et al. "China's Public Firms' Attitudes towards Environmental Protection Based on Sentiment Analysis and Random Forest Models." Sustainability 14.9 (2022): 5046.

[9] Hidayat, Tirta Hema Jaya, et al. "Sentiment analysis of Twitter data related to Rinca Island development using Doc2Vec and SVM and logistic regression as a classifier." Procedia Computer Science 197 (2022): 660-667.

[10] Li, Wei, et al. "BiERU: Bidirectional emotional recurrent unit for conversational sentiment analysis." Neurocomputing 467 (2022): 73-82.

[11] P.M. Surya Prabha, B. Subbulakshmi, Sentimental analysis using naïve bayes classifier. in International Conference on ViTECoN (2019). https://ieeexplore.ieee.org/document/8899618

[12] P. Karthika, R. Murugesari, R. Manoranjithem, Sentiment analysis of social media network using random forest algorithm. Department of Computer Science and Engineering, Kalasalingam Academy of Research and Education, (2019). https://ieeexplore.ieee.org/document/8951367

[13] Mamta, Ekbal, A., Bhattacharyya, P., Srivastava, S., Kumar, A., and Saha, T. (2020). Multi-domain tweet cor- 589 pora for sentiment analysis: Resource creation and evaluation. In LREC.

[14] Neha Nandal, (2020). Machine learning based aspect level sentiment analysis for Amazon products.
[15] Aashutosh Bhatt, Ankit Patel, Harsh Chheda, Kiran Gawande, (2015). Amazon Review Classification and Sentiment Analysis.

[16] Alexander Wallin, (2014). Sentiment analysis of Amazon reviews and perception of product features.
[17] Tanjim Ul Haque, Nudrat Nawal Saber, Faisal Muhammad Shah, (2018). Sentiment analysis on large scale Amazon product reviews.

[18] Polignano, M., Basile, P., de Gemmis, M., Semeraro, G., and Basile, V. (2019). AlBERTo: Italian BERT Language Understanding Model for NLP Challenging Tasks Based on Tweets. In Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it 2019), volume 2481. CEUR

[19] Bianchi, F., Nozza, D., and Hovy, D. (2021). FEELIT: Emotion and sentiment classification for the Italian language. In Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 76–83, Online, April. Association for Computational Linguistics.

[20] Barbieri, F., Basile, V., Croce, D., Nissim, M., Novielli, N., and Patti, V. (2016). Overview of the evalita 2016 sentiment polarity classification task. In Proceedings of third Italian conference on computational linguistics (CLiC-it 2016) & fifth evaluation campaign of natural language processing and speech tools for Italian. Final Workshop (EVALITA 2016).

[21] Malo, P., Sinha, A., Korhonen, P. J., Wallenius, J., and Takala, P. (2014). Good debt or bad debt: Detecting semantic orientations in economic texts. Journal of the Association for Information Science and Technology, 65.

[22] Takala, P., Malo, P., Sinha, A., and Ahlgren, O. (2014). Gold-standard for topic-specific sentiment analysis of economic texts. In LREC.

[23] https://www.similarweb.com/website/amazon.eg/#overview