2021 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA) | 978-1-6654-1249-0/21/$31.00 ©2021 IEEE | DOI: 10.1109/CIVEMSA52099.2021.9493576

Unmask inflated product reviews through Machine Learning

Micaela Lucia Bangerter*, Giuseppe Fenza†, Mariacristina Gallo‡, Alessia Genovese§,
Francesco David Nota¶, Claudio Stanzione‖ and Gennaro Zanfardino**
Department of Management & Innovation Systems,
University of Salerno
84084 Fisciano (SA), Italy
Email: *m.bangerter@studenti.unisa.it, †gfenza@unisa.it, ‡mgallo@unisa.it, §a.genovese28@studenti.unisa.it,
¶f.nota1@studenti.unisa.it, ‖c.stanzione8@studenti.unisa.it, **g.zanfardino3@studenti.unisa.it

***Abstract*—Product reviews play a crucial role in online customer purchase decisions. In today's marketplaces, such as Amazon, the quantity and interest of product reviews is growing, along with fierce competition between sellers. At the same time, the phenomenon of fake reviews is ever-growing. Many proposals capable of detecting fake reviews based on Machine Learning (ML) exist in literature. Nevertheless, bad practices implemented by the sellers make genuine reviews difficult to recognize. For instance, some Telegram Channels are known for giving products for free in exchange for 5-star reviews. This work focuses on the analysis of two review data streams of Amazon products. The first one is composed of the reviews corresponding to the products in the AmazonBasics category. The latter collects reviews of the products in the Telegram channels mentioned above. The analysis reveals a substantial dissimilarity between the two sources of reviews, that conducts to the construction of a ground truth dataset employed in the classification model training. The classification activity can assist during product rating interpretation, which could be invalidated by too many fake reviews. The experimental results reveal that 1&2-star reviews are good predictors of the review's trustworthiness and the product itself.**

***Index Terms*—Fake detection, Machine Learning, Product rating, trustworthiness.**

I. Introduction

Big Data represents an extensive opportunity for companies and academics. The huge amount of data generated every day could become valuable knowledge for practitioners if correctly managed. Among the *5V's* of Big Data, *Veracity* and *Value* create significant challenges; but, starting from reliable information and applying suitable analytics it is possible to produce valuable knowledge.

From the marketing point of view, the objectives of Big Data analysis regard understanding and meeting customer needs [1]. In this sense, the analysis of User-Generated Content (UGC) assumes a relevant impact. Collecting customers' interests through posts, reviews, etc., and considering the time factor related to such interests [2] could drive, for example, targeted advertising [3], and contents adaptation [4]. In this scenario, the Amazon marketplace plays an important role in Big Data analytics. Amazon's goal is to learn as much as possible about consumer shopping habits to deliver better experiences to customers themselves. This vision contributes to make Amazon one of the Big Five companies in the U.S. in the information technology industry, along with Google, Apple, Microsoft, and Facebook [5]. Data is collected from customers while they browse to build and fine-tune the recommendation engine. Consumer data empower the advertising business and enable the launch of Amazon's brand: *AmazonBasics*. Amazon sells its private-label products in its Marketplace, right alongside nearly identical products from independent sellers, and tries to consider consumers' opinions to compete better and maintain good quality standards.

Despite Amazon's best efforts, Amazon sellers have found all kinds of ways to artificially inflate their products' ratings (and lower their competitors' ratings artificially). The vast size of the platform, coupled with the ferocious competition among sellers to get higher product rankings, has generated **fake reviews** proliferation. The recruitment usually starts from Social Networks, such as Facebook and Telegram, where numerous groups can reach up to 200 thousand members ready to release fake reviews for a fee. This problem impacts customers' perception and purchase choices.

In this scenario, this work tries to evaluate reviews' integrity and provide trustworthiness to average product ratings. The idea starts from the fact that sellers often exploit social networks and messaging systems, such as Telegram, to give products for free in exchange for 5-star reviews. On the other hand, since Amazon acts in its best interest in maintaining the review system free of tampering, we consider that the AmazonBasics products have all genuine reviews. A labeled dataset used during training is constructed, starting from two data sources. The first one is the tampered collection of reviews related to products on Telegram channels and, the second, the untampered AmazonBasics ones. The objective is to train a classification model to recognize the *genuineness* of products based on their reviews by comparing sources with different trustworthiness levels [6], [7]. Most likely, products with many fake reviews have an average rating far from the reality that the customers must consider before purchasing. For this reason, after the training, the learning model should be employed to understand the level of reliability of a product's average rating based on their reviews. Therefore, the system collects reviews, analyzes them, and establishes the authenticity of the examined

978-1-6654-1249-0/21/$31.00 ©2021 IEEE

product. Regarding the adoption of the system, a manager or expert could exploit it to realize, for example, an investigation about *Company X*. He/she can start from multiple Company's products and ask the system to work on related reviews, by classifying them and make a hypothesis about its level of trustworthiness. Alternatively, the manager/expert could search for a specific product of the target Company and make the same analysis. Classification results give the expert the ability to evaluate the level of reliability of the target Company.

The paper presents a Big Data oriented framework inspired by the Lambda architecture. It works by collecting and analysing data, also in real-time. The experimentation reveals promising results in terms of classification accuracy, especially by considering 1&2-star reviews.

The remaining part of the document is structured as follows. Section II cites some related works in the area of fake review detection. Section III describes the overall workflow. Experimentation results and implementation details are reported in Section IV. Finally, Section V concludes the work and introduces some possible extensions.

II. Related Work

Interest in consumers' reviews attracts much attention among companies and research society. Numerous works are focused on designing fake review classifiers through Machine Learning [8], [9] and Deep Learning [10] approaches. A text-oriented approach, using term frequency, Latent Dirichlet Allocation (LDA) and word2vec to extract features, is presented in [11]. Beyond classic text classification approaches, recently, researchers used sentiment classification techniques [12] to distinguish between *fair* and *unfair* reviews. The authors of [13] propose a time-aware solution that searches for patterns in time windows of specific sizes and identifies outliers in product reviews.

Most of the available state-of-the-art solutions utilize supervised learning approaches. However, obtaining reliable labeled datasets of fake reviews could be very hard [14]. This work starts from this technical issue to automatically extract a ground truth about fake reviews leveraging reviews of AmazonBasics products and Telegram dedicated channels. In this sense, the literature has not developed much yet. Some approaches focus on the biasing of ratings [15]; a trustworthy service rating system based on users' reporting is adopted in [16]. Other approaches mainly consider the user's credibility during the product rating evaluation [17], [18].

III. Overall WorkFlow

The proposed framework consists of two main components. The first one aims to the **creation** of a classification model able to predict the genuineness of a product based on their reviews. It collects data from heterogeneous sources, blends, stores, and analyzes them. The objective is a suitable feature extraction for creating a labeled dataset subsequently employed for machine learning purposes. The second activity resides in the **adoption** of the created learning model to understand the level of reliability of products in Amazon.

The following sections will detail each of the aforementioned phases.

A. Creating the Classification model

The creation of the classification model consists of two sub-phases: (1) Constructing the labeled dataset employed for training the classification model; (2) Classification model training.

As shown in Figure 1, the process of labeled dataset construction starts with data collection about product reviews from two different data sources with varying reliability levels. Telegram, which hosts many channels in which sellers give products for free in exchange for 5-star reviews, is considered a tampered review source. AmazonBasics has been chosen to retrieve products that have all genuine reviews. In the first case, Telegram shares CSV files containing links to all products. A scraper (a program that gathers data) collects AmazonBasics reviews and provides a list of product links. Another type of scraper uses both lists of links to identify and extract review contents. Review contents are stored in a MongoDB database containing specific collections (i.e., genuine/dubious) for each type of review.

The whole data collection process is handled by Apache NiFi[1].

Subsequently, review contents undergo a Natural Language Processing (NLP), where stop words and punctuation are removed, and word embeddings are created. The feature extraction process constructs a labeled dataset by exploiting extracted information and results of the word embedding application on review contents. The row labeling (i.e., genuine/dubious) is done according to review origin.

Activities related to the dataset creation are handled by Apache Spark[2].

Figure 2 shows an example of labeled dataset construction starting from sample reviews coming from dubious products on Telegram and AmazonBasics products. The NLP cleans text in terms of useless words and employs stemming and lemmatization to assist the vocabulary construction made by the word embeddings technique. The result is a dataset containing relevant features and a label for each review. Let us note that the label assumes value ”dubious” if the product associated with the review is on a fake reviews Telegram channel or value ”genuine” if the product associated with the review is an AmazonBasics one.

Once the labeled dataset is created, it is employed for training the chosen learning model. The aim is to identify the authenticity of a product based on its reviews.

B. Classification model adoption

As regards the adoption of the generated classification model, real-time stream management is implemented. Apache Kafka[3] processes a stream of submitted reviews to classify products and allows suitable alerts for the user.

[1] https://nifi.apache.org/

[2] https://spark.apache.org/

[3] https://kafka.apache.org/

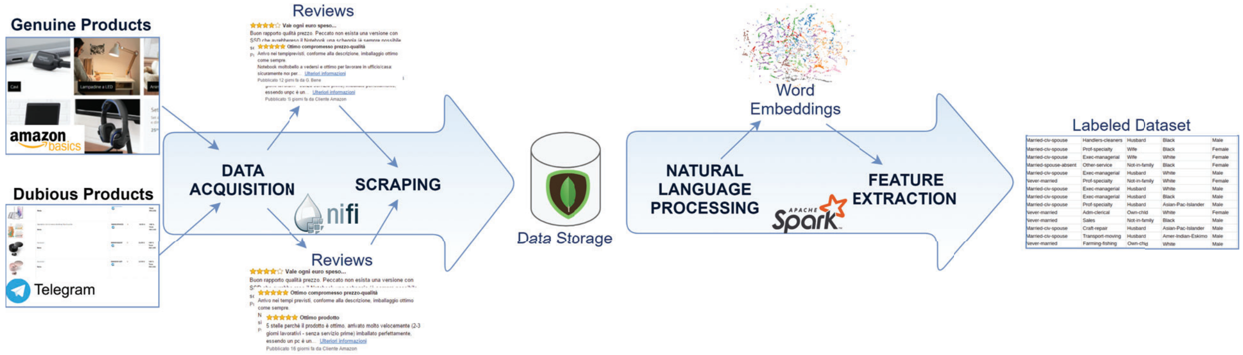
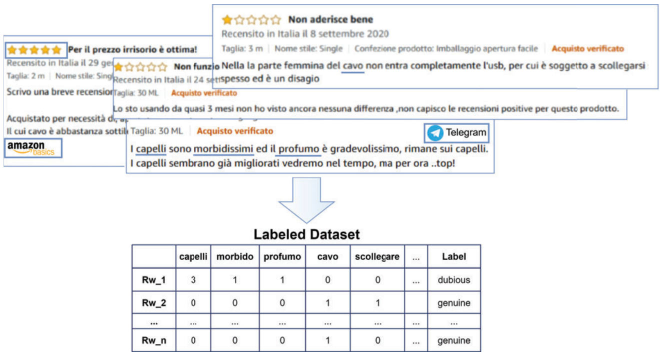


Fig. 1. Labeled dataset construction workflow



★★★★★ Per il prezzo irrisorio è ottima!
Recensito in Italia il 29 ge
Taglia: 2 m | Nome stile: Sing
Scrivo una breve recensio
Acquistato per necessità d
Il cui cavo è abbastanza sottil
amazon basics

★☆☆☆☆ Non aderisce bene
Recensito in Italia il 8 settembre 2020
Taglia: 3 m | Nome stile: Single | Confezione prodotto: Imballaggio apertura facile | Acquisto verificato
Nella la parte femmina del cavo non entra completamente l'usb, per cui è soggetto a scollegarsi spesso ed è un disagio

★☆☆☆☆ Non funzio
Recensito in Italia il 24 set
Taglia: 30 ML | Acquisto verificato
Lo sto usando da quasi 3 mesi non ho visto ancora nessuna differenza ,non capisco le recensioni positive per questo prodotto.

Taglia: 30 ML | Acquisto verificato
Telegram
I capelli sono morbidissimi ed il profumo è gradevolissimo, rimane sui capelli. I capelli sembrano già migliorati vedremo nel tempo, ma per ora ..top!

Labeled Dataset

	capelli	morbido	profumo	cavo	scollegare	...	Label
Rw_1	3	1	1	0	0	...	dubious
Rw_2	0	0	0	1	1		genuine
...	...	...	...	...	...	...	...
Rw_n	0	0	0	1	0	...	genuine

Fig. 2. Features Extraction example

Fig. 3 shows a simplified representation of the proposed Kafka architecture that consists of:

- Amazon reviews Collector. A processor that, based on the input list of products, collects concerning Amazon reviews and produces a reviews stream.
- Feature Extractor. A processor that replicates the feature extraction process adopted during the creation of the training dataset. The objective is to obtain classifiable rows transferred to the Classifier Processor through the test-set stream.
- Classifier Processor. A processor that includes the classification model created in the specific phase. It predicts the genuineness of the product related to the input review coming through the test-set stream. Dubious classified products form the warnings stream;
- Alerting Notifier. Based on the content of the warnings stream, it produces alerts for the user. This processor uses a fixed threshold to establish when generating alerts. Once the number of dubious review classifications for a product exceeds the threshold, a warning is returned.

A user (i.e., a manager or expert) interacts with the system through the "Amazon reviews Collector". He/she inputs the Amazon seller page of a target Company or the Amazon page of a specific target Company's product. Such information fills a *Product list* subsequently adopted by the "Amazon reviews Collector" to establish what reviews are reliable.

IV. Experimentation

During the experimentation, the training and testing of multiple learning models have been performed, as well as an assessment of their performances. The objective is to identify the best performing technique in terms of supervised learning to predict the genuineness of a product through its reviews. Two investigations are performed: the first one consists of training a model over 5-star reviews, and the latter consists of training a model over 1&2-star reviews. As described in the following section, these two types of reviews are the most relevant for discriminating between genuine and dubious products.

Three different classification algorithms are trained: Random Forest, Multilayer Perceptron, and K-Nearest Neighbors. The training set represents 70% of the data while the test set 30%. Through a Big Data-oriented implementation of the framework, the collection of real data and its exploitation are done. The framework, inspired by the Lambda Architecture [19], consists of the subsequent layers:

- A Batch Layer in which data is stored in the Master Dataset based on MongoDB. This layer aims to collect data about products (from Amazon and Telegram) subsequently adopted during the training.
- A Serving Layer: Apache Spark carries out the Batch Processing to produce Batch Views to perform analytics and construct the training dataset, ensuring good scalability of the framework.
- A Speed Layer: Apache Kafka, through a queue message management system, makes real-time analysis of the most recent reviews concerning specific products.

A. Dataset

The Experimentation regards a dataset of reviews organized as follows: 28544 reviews coming from Telegram products, and 13609 reviews from AmazonBasics products. Total genuine products are 158 while fake ones 1059. The considered reviews are in the Italian language.

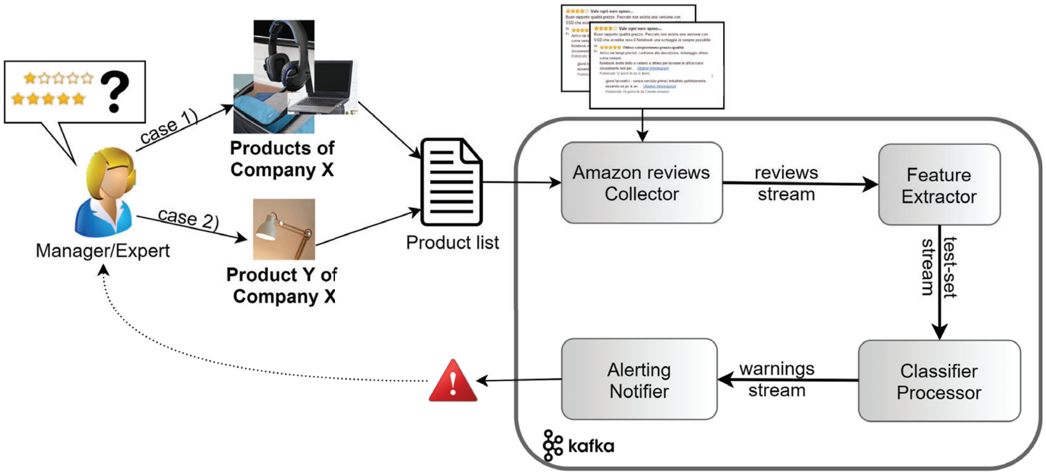


Fig. 3. Real-time streams analysis

The evaluated dataset reveals meaningful differences between reviews of products considered genuine (about AmazonBasics' products) and the dubious ones collected through Telegram. Table I reports the percentage of each star review for both, genuine and dubious products. One of the main evident results is that genuine products have a more naturally distributed rating ratio.

TABLE I
PERCENTAGE FOR EACH STAR RATING IN EACH PRODUCT CATEGORY.

Stars	Genuine Products	Dubious Products
5 stars	64.69%	81.53%
4 stars	19.46%	6.26%
3 stars	7.26%	3.62%
2 stars	3.58%	2.54%
1 star	5.01%	6.05%

By observing box plots in Fig. 4, we can see an evident difference in distributions and maximums between the average rating of genuine products on the left and the dubious ones on the right. There is also a higher variance in the distribution representing products with fake reviews. As soon as the fake review mechanism stops receiving funding, the average product rating dips, showing the low quality of products.

Another essential difference between real and fake reviews is their different text length average and distribution. As shown in Fig. 5, 5-star reviews of genuine products are longer than the 5-star ones for dubious products. It could be associated with the fact that customers want to sincerely share their positive experiences with the purchase. On the other hand, for fake reviews, users are only asked to provide a small number of sentences flattering the product to qualify for a refund. On the contrary, for 1-star reviews, by looking at AmazonBasics products, if a customer has a 1-star experience, there is surely a story to be told and shared through the use of multiple sentences. In contrast, the angry customers that found out to have bought a rip-off rated as 5-star only need a few sentences to explain their bitterness about the situation (Fig. 6).

Considering the text content of reviews, it emerges that 1-star fake reviews have a high amount of complaining words such as "*purtroppo, pessimo, soldi buttati, inutile*" (which meaning is respectively "unfortunately, bad, wasted money, useless"), and they refer to doubts on the reliability of other reviews. On the other hand, real reviews mention problems with either shipping or defective functionalities. For genuine reviews, the use of positive words is more diversified, while for fake ones, the same words are frequently repeated, suggesting a copy-paste-like behavior among reviewers.

B. Measures

Let P^* the set of genuine/dubious products, and P the set of genuine/dubious products recognized by the system. The performance of the evaluated algorithms is weighted by means of the *Accuracy* and *F-Measure* metrics, defined as follows.

$$Accuracy = \frac{T}{N} \tag{1}$$

where T is the number of correct predictions, and N is the total number of examined cases.

$$F-Measure = 2 \cdot \frac{|Precision \cdot Recall|}{|Precision + Recall|} \tag{2}$$

where:

$$Precision = \frac{|P^* \bigcap P|}{|P|} \tag{3}$$

$$Recall = \frac{|P^* \bigcap P|}{|P^*|} \tag{4}$$

C. Results & Discussion

Table II collects results about the execution of selected Machine Learning algorithms. Random Forest reaches the best results in terms of accuracy (in bold), without the need for much tuning of hyperparameters. The real implementation adopts the *RandomForestClassifier* class in the python scikit-learn library[4], for which hyperparameters are set as follows.

[4] https://scikit-learn.org/stable/index.html

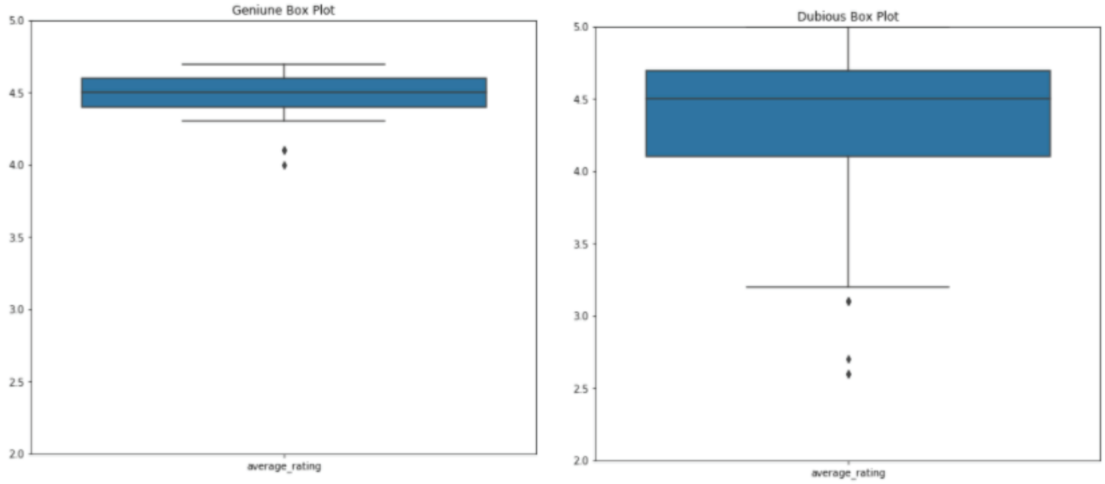


Fig. 4. Reviews distributions on two categories

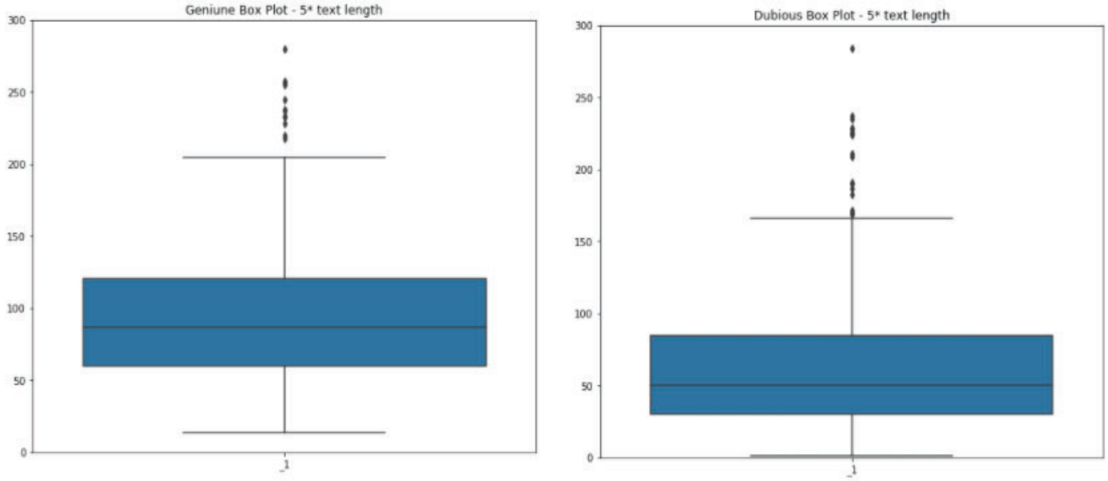


Fig. 5. 5-star reviews text length distributions on two categories

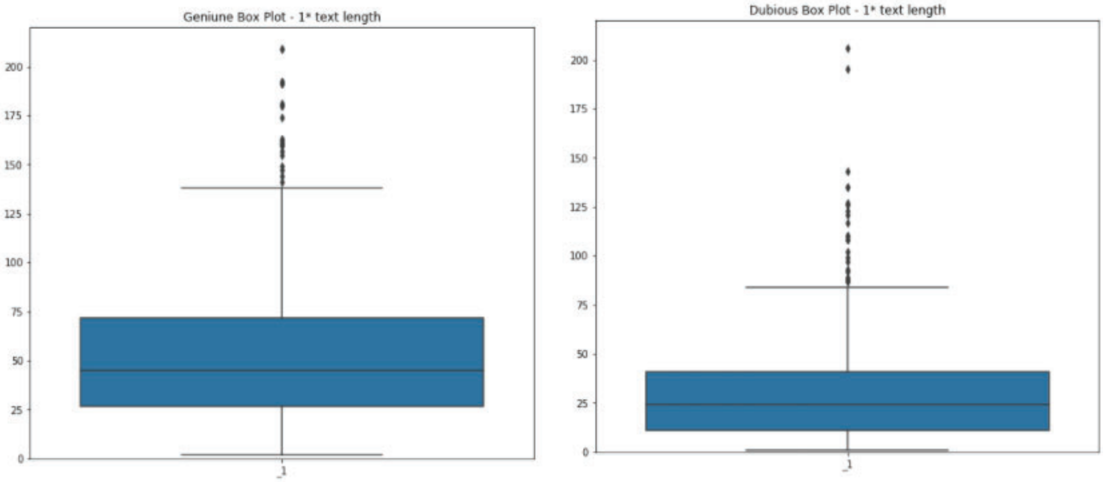


Fig. 6. 1-star reviews text length distributions on two categories

TABLE II
CLASSIFICATION RESULTS.

Technique	Stars	Classification	F-measure	Accuracy
Random Forest	5 stars	Genuine	0.73	**0.77**
		Dubious	0.80	
	1&2 stars	Genuine	0.84	**0.91**
		Dubious	0.94	
MLP	5 stars	Genuine	0.71	0.75
		Dubious	0.77	
	1&2 stars	Genuine	0.85	0.91
		Dubious	0.94	
k-NN	5 stars	Genuine	0.57	0.58
		Dubious	0.58	
	1&2 stars	Genuine	0.37	0.74
		Dubious	0.83	

n_estimators = 100, criterion = gini, min_samples_split = 2, min_samples_leaf = 1, max_features = auto, bootstrap = True. In Table II, it is also evident that 1&2-stars reviews reveal a better ability to discriminate between genuine and dubious products with respect to 5-star ones. It emerges that the low-stars reviews better help to recognize angry customers that disagree with the unusual amount of 5-stars reviews about the considered product.

V. CONCLUSION AND FUTURE WORKS

This work presents a Big Data oriented framework implementing dubious product detection based on the evaluation of their reviews. The framework produces an alert for the user regarding the low level of trustworthiness of a product. Such information helps the user to filter out dubious products or products sold by fraudulent vendors. This paper also contributes to the identification of a ground truth guiding the training set construction for Amazon reviews. Furthermore, the experimentation, of three different supervised learning algorithms, is done.

In the future, it will be necessary to enlarge the analysis by also including reviews coming from countries different from Italy. Moreover, it will be useful to define a learning model that updates itself at specific time intervals using new product reviews. In this sense, time or performance-based criteria should be defined. Finally, a potential improvement consists in the definition of a measure to adapt the average rating of a product based on classification results on a real-time reviews stream.

REFERENCES

[1] A. Kumar, R. Shankar, and N. R. Aljohani, "A big data driven framework for demand-driven forecasting with effects of marketing-mix variables," *Industrial marketing management*, vol. 90, pp. 493–507, 2020.

[2] C. De Maio, G. Fenza, M. Gallo, V. Loia, and M. Parente, "Social media marketing through time-aware collaborative filtering," *Concurrency and Computation: Practice and Experience*, vol. 30, no. 1, p. e4098, 2018.

[3] C. De Maio, M. Gallo, F. Hao, V. Loia, and E. Yang, "Fine-grained context-aware ad targeting on social media platforms," in *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2020, pp. 3059–3065.

[4] C. De Maio, G. Fenza, M. Gallo, V. Loia, and M. Parente, "Time-aware adaptive tweets ranking through deep learning," *Future Generation Computer Systems*, vol. 93, pp. 924–932, 2019.

[5] S. Galloway, *The four: the hidden DNA of Amazon, Apple, Facebook and Google*. Random House, 2017.

[6] C. De Maio, G. Fenza, M. Gallo, V. Loia, and A. Volpe, "Cross-relating heterogeneous text streams for credibility assessment," in *2020 IEEE Conference on Evolving and Adaptive Intelligent Systems (EAIS)*. IEEE, 2020, pp. 1–8.

[7] M. De Rosa, G. Fenza, A. Gallo, M. Gallo, and V. Loia, "Pharmacovigilance in the era of social media: Discovering adverse drug events cross-relating twitter and pubmed," *Future Generation Computer Systems*, vol. 114, pp. 394–402, 2021.

[8] R. Barbado, O. Araque, and C. A. Iglesias, "A framework for fake review detection in online consumer electronics retailers," *Information Processing & Management*, vol. 56, no. 4, pp. 1234–1244, 2019.

[9] E. Elmurngi and A. Gherbi, "An empirical study on detecting fake reviews using machine learning techniques," in *2017 seventh international conference on innovative computing technology (INTECH)*. IEEE, 2017, pp. 107–114.

[10] N. Jain, A. Kumar, S. Singh, C. Singh, and S. Tripathi, "Deceptive reviews detection using deep learning techniques," in *International Conference on Applications of Natural Language to Information Systems*. Springer, 2019, pp. 79–91.

[11] S. Jia, X. Zhang, X. Wang, and Y. Liu, "Fake reviews detection based on lda," in *2018 4th International Conference on Information Management (ICIM)*. IEEE, 2018, pp. 280–283.

[12] E. I. Elmurngi and A. Gherbi, "Unfair reviews detection on amazon reviews using sentiment analysis with supervised learning techniques." *J. Comput. Sci.*, vol. 14, no. 5, pp. 714–726, 2018.

[13] W. Liu, J. He, S. Han, F. Cai, Z. Yang, and N. Zhu, "A method for the detection of fake reviews based on temporal features of reviews and comments," *IEEE Engineering Management Review*, vol. 47, no. 4, pp. 67–79, 2019.

[14] A. Mukherjee, V. Venkataraman, B. Liu, N. Glance *et al.*, "Fake review detection: Classification and analysis of real and pseudo reviews," *Technical Report UIC-CS-2013–03, University of Illinois at Chicago, Tech. Rep.*, 2013.

[15] M. V. Gopalachari, "Dbt recommender: Improved trustworthiness of ratings through de-biasing tendency of users," *Int. J. Intell. Eng. Syst*, vol. 11, no. 2, pp. 85–92, 2018.

[16] J. Du, E. Gelenbe, C. Jiang, H. Zhang, Y. Ren, and H. V. Poor, "Peer prediction-based trustworthiness evaluation and trustworthy service rating in social networks," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 6, pp. 1582–1594, 2018.

[17] M. Işik and H. Dağ, "A recommender model based on trust value and time decay: Improve the quality of product rating score in e-commerce platforms," in *2017 IEEE International Conference on Big Data (Big Data)*. IEEE, 2017, pp. 1946–1955.

[18] M. Gan, "Cousin: A network-based regression model for personalized recommendations," *Decision Support Systems*, vol. 82, pp. 58–68, 2016.

[19] N. Marz and J. Warren, "Lambda architecture," 2015.