Information Processing and Management 56 (2019) 1234–1244

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

A framework for fake review detection in online consumer electronics retailers

Rodrigo Barbado, Oscar Araque*, Carlos A. Iglesias

Intelligent Systems Group, Department of Telematic Engineering Systems, Universidad Politécnica de Madrid, ETSI Telecomunicación, Avda. Complutense, 30, Madrid, 20840, Spain

ARTICLE INFO

Keywords:
Fake Review
Sentiment Analysis
Machine Learning
Data Analysis
Web Analytics

ABSTRACT

The impact of online reviews on businesses has grown significantly during last years, being crucial to determine business success in a wide array of sectors, ranging from restaurants, hotels to e-commerce. Unfortunately, some users use unethical means to improve their online reputation by writing fake reviews of their businesses or competitors. Previous research has addressed fake review detection in a number of domains, such as product or business reviews in restaurants and hotels. However, in spite of its economical interest, the domain of consumer electronics businesses has not yet been thoroughly studied. This article proposes a feature framework for detecting fake reviews that has been evaluated in the consumer electronics domain. The contributions are fourfold: (i) Construction of a dataset for classifying fake reviews in the consumer electronics domain in four different cities based on scraping techniques; (ii) definition of a feature framework for fake review detection; (iii) development of a fake review classification method based on the proposed framework and (iv) evaluation and analysis of the results for each of the cities under study. We have reached an 82% F-Score on the classification task and the Ada Boost classifier has been proven to be the best one by statistical means according to the Friedman test.

1. Introduction

Online consumer product reviews are playing an increasingly important role for customers, constituting a new type of WOM information (Chen & Xie, 2008). Recent research shows that 52% of online consumers use the Internet to search for product information, while 24% of them use the Internet to browse products before making purchases (Ha, yong Bae, & Son, 2015). As a result, online reviews has a strong impact on consumers' decision purchase in e-commerce, affecting the most relevant areas, such as travel and accommodations (Filieri & McLeay, 2014; Sotiriadis & Van Zyl, 2013), online retailers (Awad & Ragowsky, 2008), and entertainment (Chevalier & Mayzlin, 2006; Dhar & Chang, 2009; Zhu & Zhang, 2006). Moreover, online reviews of the same product can be found in multiples sources of information, which can be classified (Park, Gu, & Lee, 2012) according to the parties that host WOM information into internal WOMs, hosted by retailers (e.g. Amazon, Walmart, BestBuy, etc.) and external ones, hosted by independent product review providers (e.g. CNET, Yelp, TripAdvisor, Epinions, etc.).

Nevertheless, only credible reviews have a significant impact on consumers' purchase decision (Chakraborty & Bhat, 2018). Moreover, product category affects significantly the credibility of WOMs (Mudambi & Schuff, 2010). Consumer electronics product

* Corresponding author.

E-mail addresses: rodrigo.barbado.esteban@alumnos.upm.es (R. Barbado), o.araque@upm.es (O. Araque), carlosangel.iglesias@upm.es (C.A. Iglesias).

https://doi.org/10.1016/j.ipm.2019.03.002
Received 7 September 2017; Received in revised form 7 November 2018; Accepted 5 March 2019
Available online 16 March 2019

category is the most online reviewed (Chan & Ngai, 2011), based on a number of factors. On the one hand, consumer electronics usually require a significant investment, and the more valuable and expensive an item is, the more it is researched. According to a study (Riegner, 2007), consumer electronics are the product most influenced by online reviews, influencing the 24% of products acquired in this category, and being WOMs the second most influential source after search engines in this product category. On the other hand, consumers tend to research on consumer electronics products because these products change very frequently, with new products and updates of existing ones (Chakraborty & Bhat, 2018). Thus, consumers frequently trust on reviews to avoid making a wrong purchase decision (Park & Kim, 2008). As a result, Horrigan et al. Horrigan and Vitak (2008) report that more than 50% of consumer electronics buyers tend to consult several WOMs before making a purchase decision.

Some studies (Gu, Park, & Konana, 2012) show that retailer hosted online WOM influences enormously sales in low involvement products, such as books or CDs. However, consumers usually conduct a pre-sales research in high-involvement products, such as consumer electronics. Thus, in consumer electronics, retailer's internal WOM has a limited influence, while external WOM sources have a significant impact on the retailer's reputation and sales (Cui, Lui, & Guo, 2012). Hence, consumer electronics are more sensible to the effects of external WOMs, since they cannot easily act on them.

Since both consumers and retailers become overwhelmed by the huge number of available opinions in WOM internal and external sources, automatic natural language processing and sentiment analysis techniques have been frequently applied. Some of the most frequent application domains are review polarity classification (Poria, Cambria, & Gelbukh, 2016), review summarization (Potthast & Becker, 2010), competitive intelligence acquisition (Dey, Haque, Khurdiya, & Shroff, 2011) and reputation monitoring (Ziegler & Skubacz, 2006).

Given the importance of reviews for businesses and the difficulty of obtaining a good reputation on the Internet, several techniques have been used to improve online presence, including unethical ones. Fake reviews are one of the most popular unethical methods which are present on sites such as Yelp or TripAdvisor. However, according to Jindal and Liu (2007b), not all fake reviews are equally harmful. Fake negative reviews on good quality products are really harmful for enterprises, and along with fake positive reviews on poor quality products, result also harmful for consumers. Fake positive reviews on poor quality products are also harmful for competitors who offer average or good quality products but do not have so many reviews on them.

The goal of this article is analyzing the fake review problem in the consumer electronics field, more precisely studying Yelp businesses from four of the biggest cities of the USA. No prior research has been carried out in this concrete field, being restaurants and hotels the most previously studied cases. We want to prove that fake review detection problem in online consumer electronics retailers can be solved by machine learning means and to show if the difficulty of achieving it depends on geographic location.

In order to achieve this goal, we have followed a principled approach. Based on literature review and experimentation, a feature framework for fake review detection is proposed, which includes some contributions such as the exploitation of the social perspective. This framework, so called Fake Feature Framework (F3), helps to organize and characterize features for fake review selection. F3 considers information coming from both the user (personal profile, reviewing activity, trusting information and social interactions) and review elements (review text), establishing a framework with which categorize existing research. In order to evaluate the effectiveness of the features defined in F3, a dataset from the social Yelp in four different cities has been collected and a classification model has been developed and evaluated.

The reminder of the paper is structured as follows. Section 2 reviews the state of the art on fake review detection on other domains. Afterwards, Section 3 presents the followed methodology and also introduces the proposed feature framework. Experimentation is detailed in Section 4. Finally, Section 5 highlights and discusses the main obtained results.

2. Related work

The task of fake review detection has been studied since 2007, with the analysis of review spamming (Jindal & Liu, 2007b). In this work, the authors analyzed the case of Amazon, concluding that manually labeling fake reviews may result challenging, as fake reviewers could carefully craft their reviews in order to make them more reliable for other users. Consequently, they proposed the use of duplicates or nearly-duplicates as spam in order to develop a model that detects fake reviews (Jindal & Liu, 2007b). Research on distributional footprints has also been carried out, showing a connection between distribution anomalies and deceptive reviews from Amazon products and TripAdvisor hotels (Feng, Xing, Gogar, & Choi, 2012).

Fake review detection is a specific application of the general problem of deception detection, where both verbal and nonverbal clues can be used (Fitzpatrick, Bachenko, & Fornaciari, 2015). Fake review detection research has mainly exploited textual and behavioral features, while other approaches have taken into account social or temporal aspects.

Textual features have been proposed in several papers. Ott, Choi, Cardie, and Hancock (2011) employed psycholinguistic features based on LIWC (Tausczik & Pennebaker, 2010) combined with standard word and Part of Speech (POS) n-gram features. Mukherjee, Venkataraman, Liu, and Glance (2013) extend that work including also style and POS based features, such as deep syntax and POS sequence patterns. However, the detection of fake reviews based only on textual features is challenging. Other articles propose additional textual features such as semantic similarity and emotion Li, Feng, and Zhang (2016), a wide variety of lexical and syntactic features (Dewang & Singh, 2015) and deeper details such as understandability, level of details, writing style and cognition indicators (Banerjee, Chua, & Kim, 2015).

Behavioral features refer to nonverbal characteristics of review activity, such as the number of reviews or the time and device where the review was posted. They were used in order to improve the classification model resulting in encouraging results. Jindal and Liu (2008) introduced behavioral features on Amazon reviews, distinguishing among review features (e.g. number of feedbacks, position of the review, textual features, rating features, etc.), product features (e.g. price, sales rank) and reviewer features (e.g.

Table 1
Reviewed works classified according to F3 framework.

Reference	Year	Domain	Algorithms	Personal	Social	Review Activity	Trust	Review centric
Liu et al. (2017)	2017	E-Commerce	–		x	x	x	x
Zhang et al. (2016)	2016	Hotel	SVM, NB, DT, RF, LR	x	x	x	x	x
Li et al. (2016)	2016	Restaurant	SVM, NB, DT		x	x	x	x
Heydari et al. (2016)	2016	E-Commerce	–		x	x		
Luca and Zervas (2016)	2016	Restaurant	–	x	x	x		x
Dewang and Singh (2015)	2015	Hotel	NB					x
Li et al. (2015)	2015	Restaurant	SVM			x		x
Banerjee et al. (2015)	2015	Hotel	SVM, NB, RF, DT, LR					x
Fusilier et al. (2015)	2015	Hotel	SVM, NB					x
Fornaciari and Poesio (2014)	2014	E-Commerce	SVM				x	
Akoglu et al. (2013)	2013	Software	–				x	
Mukherjee et al. (2013)	2013	Hotel, Restaurant	SVM			x	x	x
Fei et al. (2013)	2013	Software	SVM			x	x	
Li et al. (2011)	2011	E-Commerce	NB	x	x	x	x	x
Ott et al. (2011)	2011	Hotels	SVM, NB					x
Wang et al. (2011)	2011	E-Commerce	–				x	
Jindal and Liu (2008)	2008	E-Commerce	SVM			x	x	x

average rating, ratio of the number of reviews that the reviewer wrote which were the first reviews, etc.). In another work, Zhang, Zhou, Kehoe, and Kilic (2016) explore the effect of both textual and behavioural features in the restaurant and hotel domain, showing that non-textual features result more relevant for the task of fake review detection. Also, regarding the restaurant domain, some interesting findings were described by Luca and Zervas (2016). Restaurants are more likely to make review fraud when they have a lower reputation, including having few reviewers or bad scorings.

Apart from using textual and behavioral features, other methodologies were followed for the fake review detection task. Wang, Xie, Liu, and Yu (2011) proposed a review graph with the aim of capturing relationships between reviewers, reviews and stores reviewed by the reviewers. Making use of this graph, an iterative model was used to identify suspicious reviewers. Following also a graph model, network effects were analyzed by Akoglu, Chandy, and Faloutsos (2013), following two steps: User and review scoring for fraud detection and grouping for visualization.

Another methodological approach focuses on temporal aspects, and concerns the burstiness of reviews and their impact on businesses. Bursts of reviews can be either due to sudden popularity of products or spam attacks (Fei et al., 2013), which were also analyzed in (Liu, Xu, Ai, & Wang, 2017) along with other behavioral and textual features. A deeper time series approach was made by Heydari, Tavakoli, and Salim (2016) and Li et al. (2016) propose other types of features such as review density in temporal windows, along with semantic and emotion features. Spatial and temporal features were used in a Chinese site by Li, Chen, Mukherjee, Liu, and Shao (2015).

Regarding classification algorithms, Support Vector Machine (Vapnik, Golowich, & Smola, 1997) was the most used one followed by Naive Bayes (Friedman, Geiger, & Goldszmidt, 1997), Decision Tree (Breiman, Friedman, Stone, & Olshen, 1984), Random Forest (Breiman, 2001) and Logistic Regression (Cox, 1958) as shown in Table 1.

Apart from supervised learning, other approaches have been followed, since collecting data for experiments is a hard task. In Li, Huang, Yang, and Zhu (2011), authors propose a prediction model based on semi-supervised learning and a set of textual and behavioural features. Additionally, Fusilier, y Gómez, Rosso, and Cabrera (2015) propose a semi-supervised technique called PU-learning.

The described prior research highlighted in this section has been organized according to our proposed framework as shown in Table 1. Regarding the field of application, this table shows that previous research has been centered around restaurant and hotel domains when considering reviews about businesses. To the extent of our knowledge, the study of the problem of fake review detection on the consumer electronics domain is a novel work. Moreover, its application is highly relevant given that consumer electronics are more sensible to an unethical WOM misuse.

3. Methodology

The methodology followed in this article is shown in Fig. 1. The first step is building the dataset from Yelp by web scraping means (Section 3.1). Then, a feature model is defined and computed (Section 3.2) for training a classifier that detects fake reviews (Section 4).

3.1. Scraping process

As the consumer electronics field has not been studied before, there is not an available dataset to experiment with, so the starting point consists on data collection. Furthermore, Yelp's filter has been used as a reference for labelling reviews as fake or not since, according to its CEO, Yelp's filtering algorithm has evolved over the years to filter fake reviews (Mukherjee et al., 2013). Also, this filter has been claimed to be highly accurate (Weise, 2011).

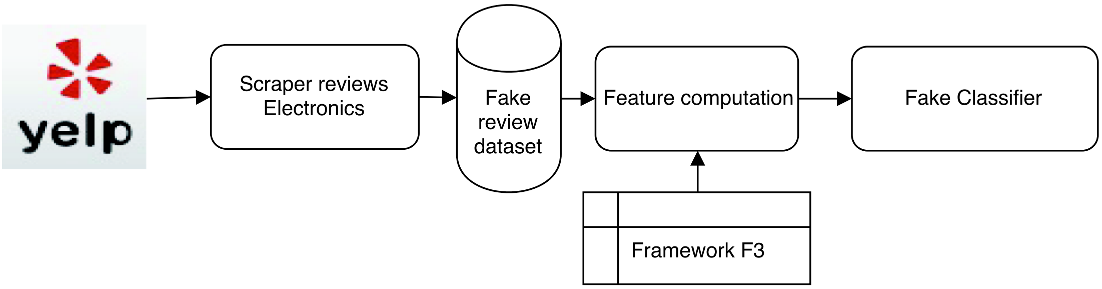

Fig. 1. Methodology.

As Yelp shows both trustful and filtered reviews, it is possible not only to extract information about reviews and users but also whether if they had been filtered or not. So, the first step was developing a web scraper to gather the necessary information for the experiments. This web scraper was programmed in Python using Scrapy (Kouzis-Loukas, 2016), a library which offers the possibility of building web crawlers. The resulting corpus was compound by reviews from four important cities of the USA: New York, San Francisco, Los Angeles and Miami. Yelp offers the possibility of searching category businesses in each city, so the scraping process is focused on the pages in which all electronic businesses from the different selected cities appeared. For each of these pages, all businesses appearing on it were selected in order to retrieve their information.

Secondly, for each business site all the reviews appearing on it were collected along with their user profile URL. In this step, reviews were also labeled as fake or not according to the Yelp filter.

The last step of the corpus creation consisted on gathering all the information appearing on each user profile site through the URLs which were collected in the previous step. Each of those profile sites show all the necessary information to develop the features of the proposed framework.

As a result of the scraping process, the dataset shown in Table 2 has been extracted. As expected, the distribution of fake and trustful reviews is highly unbalanced. In Mukherjee et al. (2013) it was shown that around 14% of reviews were fake. For our experiments, we crawled the same amount of reviews for each category, thus obtaining a balanced dataset. The dataset contains labeled reviews but also social aspects related to the user profile and her networking activity.

An exploratory data analysis of the dataset reveals interesting information. Table 3 shows the most frequent words, which are similar between all cities. Moreover, and what is more important for the fake review detection task, there are not significant differences of popular words among trustful and fake reviews. This fact indicates that text features could not be very relevant for fake review discrimination.

On the other hand, the user information obtained from our web scraper shows clear unequal statistical values between fake and trustful users. This information is presented in Table 4, showing the mean value, standard deviation and maximum values for each of the fields extracted from the user's profile site from Yelp (minimum values were not included as they were always zero). The maximum values are shown for informative purposes, as they were afterwards normalized. It can be observed that fake reviewers tend to give lower ratings on their reviews than trustful reviewers, being their mean values 1.1 stars and 2.79 stars respectively from a maximum of 5.

3.2. Fake Feature Framework (F3)

In this section we introduce a feature framework Fake Feature Framework (F3) for organizing the extraction and characterization of features in fake detection. Its definition is inspired on the analysis of previous research, and includes a novel definition of social features. Previous works have classified features into textual, behavioural and product features (Jindal & Liu, 2008). Our main contribution consists in providing a more detailed classification of user centric features, taking advantage of the social aspects of a social network such as Yelp.

As shown in Fig. 2, our framework distinguishes review centric and user centric features.

User centric features consider information related to how users behave in a social network such as Yelp and which information users provide; and are divided into four types: **Personal Features (P), Social Features (S), Reviewing Activity Features (RA)**, and **Trusting Features (T).**

The first type, **P**, is information related to user profile, such as the self description written by the user; users' businesses subscriptions, known as bookmarks; lists containing several bookmarks; registration date; updates made on self reviews; and the user's

Table 2
Sizes of reviews per city of the corpus.

City	Trustful reviews	Fake reviews
New York	2472	2472
Los Angeles	3776	3776
Miami	1409	1409
San Francisco	1799	1799

Table 3
Five most frequent words by city and class.

City	Trust	Fake
New York	Phone	Phone
	Store	Service
	Service	Store
	Screen	Back
	Time	Customer
Los Angeles	Service	Service
	Phone	Phone
	Store	Store
	Great	Great
	Place	Customer
Miami	Phone	Phone
	Service	Service
	Store	Store
	Customer	Great
	Screen	Customer
San Francisco	Phone	Phone
	Service	Service
	Time	Store
	Store	Customer
	One	Time

Table 4
Selected features distributions over the whole dataset.

Subset	Feature	Trust			Fake		
		Mean	Std	Max	Mean	Std	Max
	Profile description	0.19	0.39	1.0	0.06	0.24	1.0
Personal	Bookmark lists	36.47	183.19	5842.0	2.09	27.74	1717.0
	Lists	1.45	15.67	712.0	0.04	0.58	30.0
	Review updates	4.22	20.80	562.0	0.34	2.52	85.0
	Friends’ no. of friends	231.75	417.09	5000.0	66.77	269.39	13699.0
	Friends’ no. of reviews	80.6	189.99	2603.0	26.70	121.96	2885.0
	Profile has photo	0.76	0.43	1.0	0.41	0.49	1.0
Social	No. of followers	6.18	45.34	1782.0	0.38	5.02	263.0
	No. of friends	70.86	260.70	5000.0	13.90	106.14	5000.0
	No. of votes ‘cool’	155.91	1169.02	35842.0	5.41	112.61	5440.0
	No. of votes ‘useful’	231.35	1449.30	51012.0	8.58	128.43	6170.0
	No. of votes ‘funny’	136.18	1010.25	32844.0	4.35	92.27	4184.0
	No. of reviews	77.71	328.41	11225.0	7.78	42.14	1404.0
	Rating distribution (% 5 stars)	0.37	0.31	1	0.14	0.27	1.0
	Rating distribution (% 4 stars)	0.13	0.16	0.83	0.05	0.13	1
Review Activity	Rating distribution (% 3 stars)	0.06	0.09	1.0	0.02	0.07	0.8
	Rating distribution (% 2 stars)	0.05	0.08	0.8	0.02	0.06	0.6
	Rating distribution (% 1 star)	0.12	0.17	1.0	0.07	0.18	1.0
	Average rating	2.79	1.78	5.0	1.1	1.74	5.0
Trust	No. of photos	127.39	1135.01	57761.0	5.60	141.04	7599.0
	No. of tips	24.29	269.99	16364.0	1.27	18.56	1040.0

real name. This information is available in social review networks, such as Yelp, and can be automatically obtained.

The second type, **S**, is the feature set related to the way the user interacts with other users. In our case, we have identified these features by inspecting the Yelp social features. Our hypothesis is that social activity can help in the classification task, since social features can help to extend the context of the linguistics features. Several works have also proposed this approach in a number of applications, such as sentiment analysis (Tan et al., 2011) or stance detection (Lai, Tambuscio, Patti, Ruffo, & Rosso, 2017). Social features included here are number of friends, metrics gathered from friends such as their number of reviews or friends, number of followers, number of compliments, rank of popularity and the presence of a profile photo.

The third type, **RA** is the feature set related to the way the user posts his reviews. Some examples of reviewing activity features are review counts (Mukherjee et al., 2013), self review deviations (Jindal & Liu, 2007a), number of received votes (in Yelp there are three types: ‘cool’, ‘funny’ and ‘useful’), number of posted tips (which are shorter reviews that appear on Yelp), maximum number of reviews in a date or specific review counts in temporal windows. Additionally, metrics related to the ratios of positive and negative reviews are included here.

A last type, **T**, is the feature set that aims at pointing out inconsistencies or abnormal behaviour in the user review activity. For example, a high content similarity of the reviews of a given user could reveal that said user may be using templates for reviewing

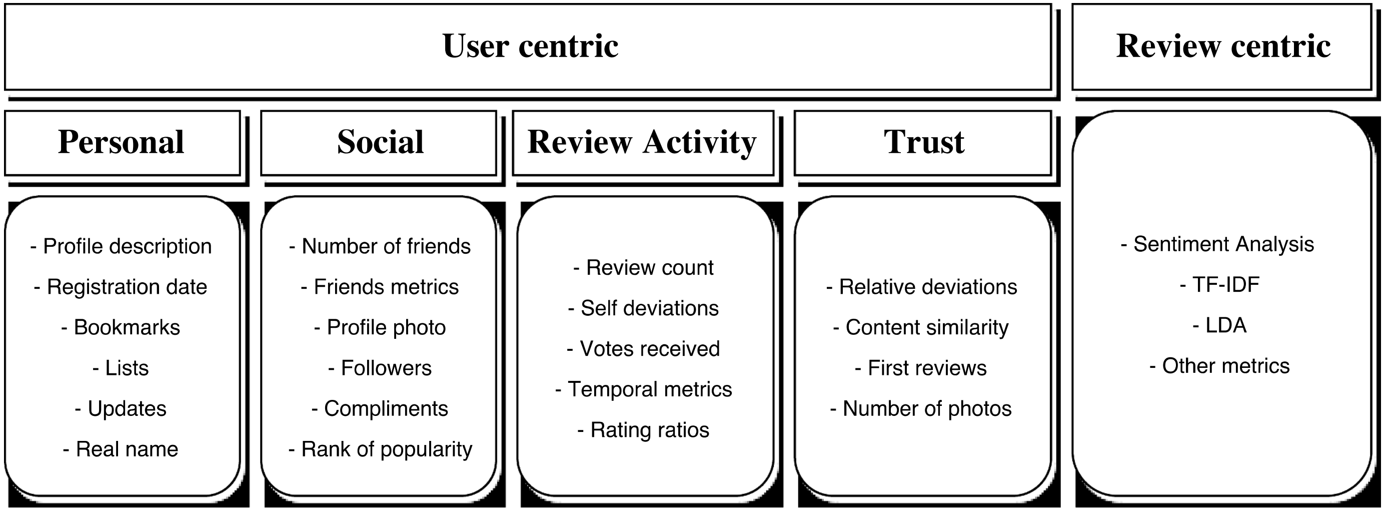

Fig. 2. Overview of the Fake Feature Framework (F3).

(Mukherjee et al., 2013). In this regard, rating deviations with respect to other users reviewing the same businesses (Jindal & Liu, 2007a) can also offer greater insight. Other features included in this type are the number of first reviews and number of uploaded photos.

In the case of **review centric features**, apart from textual metrics such as average lengths or using techniques such as Term Frequency - Inverse document frequency (TF-IDF) (Ouyang, Li, Li, & Lu, 2011), we have also included the use of **POS** tags (Petz et al., 2014), Latent Dirichlet Allocation (LDA) models (Chen, 2017), Word2Vec models (Goldberg & Levy, 2014) and sentiment analysis for enriching the available features using the information we have. Emotion analysis has also been used in prior research as well as other lexical and syntactic features.

This framework has been used for classifying state-of-the-art features as shown in Table 1.

After experimenting with different potential interesting features, the features shown in Table 4 have been selected.

4. Experimentation and evaluation

This section describes the experiments carried out to develop and evaluate a fake review classifier model based on the framework previously described. The classifier has been trained and tested over the consumer electronics dataset previously scrapped and evaluated using ten fold cross-validation. Also, we tackle the impact the user and review centric features may have on the performance, and the possible differences across different cities. After analyzing the results obtained, a statistical evaluation has been carried out.

4.1. Review centric features

The technique which best performed for analyzing review centric textual features was TF-IDF with bigrams, but it reached an F-Score below 60%. Previous research regarding the restaurant domain reached similar conclusions, stating that the text of the reviews were not a good indicator of reviews veracity (Mukherjee et al., 2013). Nevertheless, the results obtained in the particular case of consumer electronics show that text information is not useful for fake classification in this domain. In this sense, bigram representations are considered as a fairly strong baseline (Joulin, Grave, Bojanowski, & Mikolov, 2016; Wang & Manning, 2012). Obtaining a low performance with such baseline enforces the idea that the text does not serve as indicator for fake reviews. Apart from using TF-IDF, we experimented with other techniques such as LDA models or sentiment analysis, but they did not work well. The results did not improve either with the use of Word2Vec models, as these performed similar to random guessing.

4.2. User centric features

In the case of user centric features, results were clearly improved in comparison with review centric features. Regarding the experiments, we highlight the contributions of the different subdivisions of user centric features we defined in the proposed framework, which were Social Features (S), Personal Features (P), Trusting Features (T) and Reviewing Activity Features (RA). Table 5 shows the F-Scores obtained for each of the experiments done in every city, structuring the results based on city, type of features employed and classification algorithm (Logistic Regression, Decision Tree, Random Forest, Gaussian Naive Bayes, AdaBoost). Additionally, results for the combination of all cities are shown. Fig. 3 shows the F-Scores for several combinations of the proposed features aggregated by city, as obtained by AdaBoost, which is the classifier with the best performance. This way, it is easy to analyze the results considering several aspects and the following conclusions were drawn.

When comparing the impact of the different defined features, we can observe that RA features are the most relevant and S are the least ones. Nonetheless, our experiments show that social features improve the accuracy of the rest of feature types. Moreover, each combination of feature types consistently increases the performance. Thus, our initial hypothesis considering that social features could be effective is supported.

Finally, analyzing the results across the four different cities, it can be extracted that F-Scores are quite similar between New York, Miami and San Francisco, but Los Angeles achieves worse results if compared to the same pair of classifier-feature subset of any other city. When taking into account all cities at the same time, it can be seen that the better results for individual cities are not surpassed.

Table 5
F-Score results.

City	Features	LR	DT	RF	GNB	AB
All cities	S + P + T + RA	0.77	0.80	0.80	0.71	**0.81**
New York	S + P + T + RA	0.79	0.81	**0.82**	0.72	**0.82**
Los Angeles	S + P + T + RA	0.73	0.73	0.78	0.69	**0.79**
Miami	S + P + T + RA	0.78	0.81	0.81	0.71	**0.82**
San Francisco	S + P + T + RA	0.78	0.81	0.81	0.69	**0.82**
Friedman rank		3.87	2.87	2.12	5	1.12

Features legend: S (Social), P (Profile), T (Trust), RA (Reviewing Activity).
Classifier legend: LR (Logistic Regression), DT (Decision Tree), RF (Random Forest), GNB (Gaussian Naive Bayes), AB (AdaBoost).

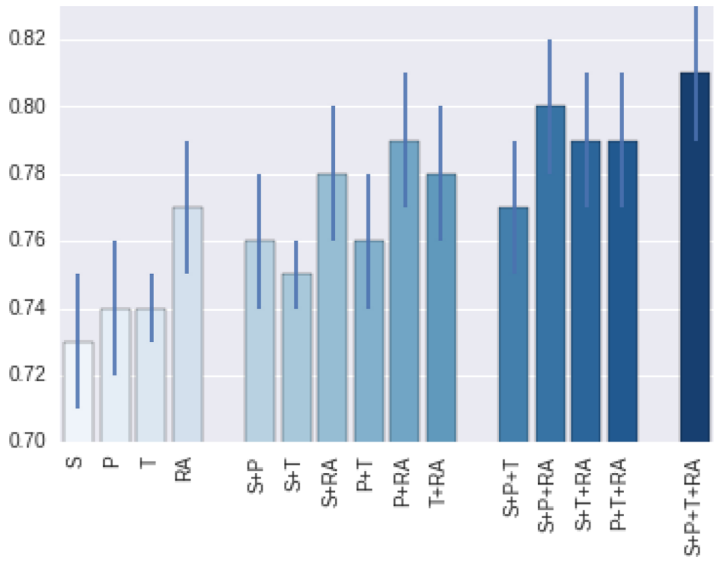

Fig. 3. Performance of review centric features using an AdaBoost classifier.

4.3. Combining review and user centric features

From the previous results, it can be seen that there is a clear difference between using review or user centric features. However, the last step was an attempt of combining both kinds of features trying to improve the model. These experiments were not satisfactory in the case of textual models which incorporate large numbers of features such as TF-IDF or Word2Vec, as those features were almost seen as noise without having any impact on the decisions taken by classifiers. Due to this, we also tried to incorporate a reduced subset of features extracted from those models but the results were the same.

In the case of adding sentiment analysis, POS features or other textual metrics to the user centric subset of features the results did neither improve as those features were also irrelevant for the classifiers. The last experiment consisted in combining the outputs of separate models built with user or review centric features following ensemble techniques. Nevertheless, it was not effective in our dataset. Some of the potential reasons of this are that fake reviewers have become more sophisticated, as well as that Yelp filter is exploiting mainly user centric features instead of linguistic ones based on their effectiveness as pointed out by Mukherjee et al. (2013). Nevertheless, other research works (Mukherjee et al., 2013) have reported a slightly improvement with the combination of bigrams and behavioural features in the restaurants and hotels domain. This discrepancy can be due to the fact the high competition in these domains and the economic incentives of fake reviews (Luca & Zervas, 2016; Mayzlin, Dover, & Chevalier, 2014).

The results of these experiments, that aim to exploit the information contained in the textual reviews, are made public online for the interested reader[1].

4.4. Statistical analysis

In order to compare the different learning models used in this work, a statistical test has been applied on the experimental data. In particular, we have chosen the Friedman test (Demšar, 2006), as this test is oriented to the comparison of several classifier methods on multiple datasets.

The Friedman test is based in a rank of each classification method in each dataset, where the best performing algorithm is assigned the rank of 1, the second best is assigned rank 2, etc. Ties in this rank are resolved by the average of their ranks. r_j^i is the rank of the j-th classification methods on the i-th dataset, where $j \in \{1, 2, \cdots, k\}$ and $i \in \{1, 2, \cdots, N\}$. With the Friedman test, the average of the ranks $R_j = \frac{1}{N}\sum_i r_i^j$ is compared. The null-hypothesis is defined as the situation where all the classification algorithms are equal, and so are their ranks R_j. Then, the Friedman statistic with $k - 1$ degrees of freedom is written as follows:

$$\chi_F^2 = \frac{12N}{k(k+1)}\left(\sum_j R_j^2 - \frac{k(k+1)^2}{4}\right)$$

Nonetheless, it is shown that there is a more useful statistic that is distributed according to the F-distribution, and has $k - 1$ and $(k - 1)(N - 1)$ degrees of freedom (Demšar, 2006). This statistic is referred to as the Friedman F, and is expressed as:

[1] http://gsi.upm.es/∼oaraque/FakeReviews/additionalmaterial.pdf.

$$F_F = \frac{(N-1)\chi_F^2}{N(k-1)-\chi_F^2}$$

If the null-hypothesis of the Friedman test is rejected, post-hoc tests can be conducted, complementing the statistical analysis. In this work, we have conducted the Nemenyi and Holm tests, with the aim of gaining insight of the differences between the analyzed classifiers.

In the Nemenyi test, all classifiers are compared to each other. In this way, the performance of a two classifiers is significantly different if their ranks differ, at least, the *critical difference*. The critical difference is computed with the following expression:

$$CD = q_\alpha \sqrt{\frac{k(k+1)}{6N}}$$

For the Holm test, the z value is computed, which allows to obtain a probability in from the normal distribution. The z value that compares the i-th and j-th classifiers is computed as:

$$z = (R_i - R_j)/\sqrt{\frac{k(k+1)}{6N}}$$

The Holm test compares the corresponding p-values of each z value with a value $\alpha/(k-1)$ value in a step-down manner. If the p-value has a lower value than the modified α value, the null-hypothesis is rejected. As soon as a certain null-hypothesis is rejected, all the remaining hypothesis are retained, that is, they can not be rejected.

The computation of the tests has been made as follows. In relation to the **Friedman** test, the ranks have been computed. The average ranks (R_i) are shown in Table 5. For all the calculations, the α value is set to 0.05. Attending to those R_i values, $\chi_F^2 = 14.5$, $F_F = 29.0$, and the critical value $F(k-1, (k-1)(N-1)) = 3.26$. Given that $F_F > F(4, 12)$, the null-hypothesis is rejected, that is, not all the classifiers have similar performance, and post-hoc tests can be conducted.

The **Nemenyi** test is then computed, selecting $q_\alpha = 2.728$, as indicated in (Demšar, 2006). With these values, the corresponding critical difference is $CD = 3.05$. Given the R_i values obtained in the Friedman test, it can be observed that the Nemenyi test points that the Gaussian Naive Bayes and the AdaBoost classifiers are significally different. Finally, the **Holm** test rejects the null-hypothesis for the AdaBoost and Random Forest classifiers, with the following values: $p|_{\mathrm{AB}} = 0.0005$, $\frac{\alpha}{i}|_{\mathrm{AB}} = 0.0125$, $p|_{\mathrm{RF}} = 0.01$ and $\frac{\alpha}{i}|_{\mathrm{RF}} = 0.017$.

In conclusion, the statistical tests point that, between all the classifiers analyzed in this work, the AdaBoost and Random Forest have the best performances in the datasets. Attending to the Friedman ranks, we highlight the AB performance in the studied problem.

5. Conclusions

In this paper we have addressed fake review detection in the consumer electronics domain. We have proposed a feature framework oriented to analysis of social sites, and we have also developed a dataset which is made available[2] for future research. Our framework is composed of two main types of features: Review centric and user centric. Review centric features are only related to the text of the review. On the other hand, user centric features show how the user behaves within the site, and are subdivided into four groups: Personal, social, reviewing activity and trust.

As shown in the article, detecting fake reviews by just reading the reviews is a challenging task for both humans and computers, since textual reviews do not usually provide fake signals to be detected. In contrast, features related to the user have been shown to be more effective. The most relevant ones are those coming from the reviewing activity. This is not surprising, since they can provide signals of abnormal reviews. The rest of features have also proven to have discrimination capability and every combination of feature types have improved the overall accuracy.

One of the conclusions of this paper is that fake reviews on the consumer electronics domain can be detected with a reasonably high F-Score using the features in the proposed F3 framework, reaching a maximum result of 82% when using Random Forest or Ada Boost classifiers. Regarding the two main kinds of features described in the F3 framework, user centric features provide clearly better results than review centric features. This last method lead to F-Scores under 60%, despite applying techniques such as TF-IDF or neural learning models such as Word2vec. This fact had also been studied in other fields such as restaurant or hotel reviews (Mukherjee et al., 2013), and it is reinforced by the difficulty of determining whether a review is fake or not by a human who reads it. Several experiments with the aim of combining both kinds of features were carried out including ensemble methods and dimensionality reduction techniques, but the results did not improve the user centric features benchmark.

The use of social features, a subset of user centric features inside F3, had not been studied before in the fake review detection problem of any field. Interestingly, this set of features has brought great results reaching a maximum of 82% F-Score in New York. As it has been said before, the most relevant features were the ones related to reviewing activity, but better results were achieved when combining different subsets of user centric features as each combination between the four subsets resulted into an increase of the F-Score. This fact indicates that the idea of dividing user centric reviews into four subsets has been successful, as features between subsets are independent and a classification system can take advantage of them when combined.

Our dataset was compound of reviews from four of the most important USA cities, and results were quite similar between all of

[2] To obtain the dataset for research purposes, please write to *o.araque@upm.es*.

them. However, F-Scores in Los Angeles were a bit lower than in New York, Miami or San Francisco. With the final results obtained from each classifier incorporating all the features, we made a statistical analysis following the Friedman test, which shows that Ada Boost performs statistically better than the rest of used classifiers.

Insights from this study can push forward the field of fake detection in social platforms. The proposed framework is a first attempt to classify and organize the features used in this emerging research area. As reported in this article, we have experimentally shown that using only the text of a review is not an effective approach, as other researchers have also shown before. In light of this, common sense suggests that fake reviewers perform a fairly good job at disguising such invalid reviews, and even machine learning methods can be tricked. Nevertheless, fake reviewers cannot hide their social network footprints and this can be a path for detecting them. The proliferation of fake reviews and more recently fake news is a social problem that is overwhelming our society, so it is needed further research. As future work, we propose to extend the proposed Fake Feature Framework (F3) on a general domain out of the e-commerce field. Additionally, this framework can also be used in other tasks, such as toxic user activity detection or as previously mentioned for fake news detection, which could involve complex semantic analysis.

Acknowledgments

This work is supported by the Spanish Ministry of Economy and Competitiveness under the R&D projects SEMOLA (TEC2015-68284-R) and EmoSpaces (RTC-2016-5053-7), by the Regional Government of Madrid through the project MOSI-AGIL-CM (grant P2013/ICE-3019), the European Union under Trivalent (2020 RIA Action Grant No. 740934 under the call SEC-06-FCT-2016), and Spanish MINETAD (TSI-102600-2016-1).

References

Akoglu, L., Chandy, R., & Faloutsos, C. (2013). Opinion fraud detection in online reviews by network effects. *ICWSM, 13*, 2–11.

Awad, N. F., & Ragowsky, A. (2008). Establishing trust in electronic commerce through online word of mouth: An examination across genders. *Journal of Management Information Systems, 24*(4), 101–121.

Banerjee, S., Chua, A. Y., & Kim, J.-J. (2015). *Using supervised learning to classify authentic and fake online reviews. Proceedings of the 9th international conference on ubiquitous information management and communication.* ACM88.

Breiman, L. (2001). Random forests. *Machine learning, 45*(1), 5–32.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees.* CRC press.

Chakraborty, U., & Bhat, S. (2018). The effects of credible online reviews on brand equity dimensions and its consequence on consumer behavior. *Journal of Promotion Management, 24*(1), 57–82.

Chan, Y. Y., & Ngai, E. W. (2011). Conceptualising electronic word of mouth activity: An input-process-output perspective. *Marketing Intelligence & Planning, 29*(5), 488–516.

Chen, L.-C. (2017). An effective lda-based time topic model to improve blog search performance. *Information Processing & Management, 53*(6), 1299–1319. https://doi.org/10.1016/j.ipm.2017.08.001.

Chen, Y., & Xie, J. (2008). Online consumer review: Word-of-mouth as a new element of marketing communication mix. *Management Science, 54*(3), 477–491.

Chevalier, J. A., & Mayzlin, D. (2006). The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research, 43*(3), 345–354.

Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society. Series B (Methodological),* 215–242.

Cui, G., Lui, H.-K., & Guo, X. (2012). The effect of online consumer reviews on new product sales. *International Journal of Electronic Commerce, 17*(1), 39–58.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research, 7*(Jan), 1–30.

Dewang, R. K., & Singh, A. (2015). *Identification of fake reviews using new set of lexical and syntactic features. Proceedings of the sixth international conference on computer and communication technology 2015.* ACM115–119.

Dey, L., Haque, S. M., Khurdiya, A., & Shroff, G. (2011). *Acquiring competitive intelligence from social media. Proceedings of the 2011 joint workshop on multilingual OCR and analytics for noisy unstructured text data.* ACM3.

Dhar, V., & Chang, E. A. (2009). Does chatter matter? the impact of user-generated content on music sales. *Journal of Interactive Marketing, 23*(4), 300–307.

Fei, G., Mukherjee, A., Liu, B., Hsu, M., Castellanos, M., & Ghosh, R. (2013). Exploiting burstiness in reviews for review spammer detection. *ICWSM, 13*, 175–184.

Feng, S., Xing, L., Gogar, A., & Choi, Y. (2012). Distributional footprints of deceptive product reviews. *ICWSM, 12*, 98–105.

Filieri, R., & McLeay, F. (2014). E-WOM and accommodation: An analysis of the factors that influence travelers' adoption of information from online reviews. *Journal of Travel Research, 53*(1), 44–57.

Fitzpatrick, E., Bachenko, J., & Fornaciari, T. (2015). Automatic detection of verbal deception. *Synthesis Lectures on Human Language Technologies, 8*(3), 1–119.

Fornaciari, T., & Poesio, M. (2014). *Identifying fake amazon reviews as learning from crowds. Proceedings of the 14th conference of the european chapter of the association for computational linguistics*279–287.

Friedman, N., Geiger, D., & Goldszmidt, M. (1997). Bayesian network classifiers. *Machine Learning, 29*(2-3), 131–163.

Fusilier, D. H., y Gómez, M. M., Rosso, P., & Cabrera, R. G. (2015). Detecting positive and negative deceptive opinions using pu-learning. *Information Processing & Management, 51*(4), 433–443. https://doi.org/10.1016/j.ipm.2014.11.001.

Goldberg, Y., & Levy, O. (2014). Word2vec explained: Deriving Mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722*.

Gu, B., Park, J., & Konana, P. (2012). Research note-the impact of external word-of-mouth sources on retailer sales of high-involvement products. *Information Systems Research, 23*(1), 182–196. https://doi.org/10.1287/isre.1100.0343.

Ha, S. H., yong Bae, S., & Son, L. K. (2015). Impact of online consumer reviews on product sales: Quantitative analysis of the source effect. *Applied Mathematics, 9*(2L), 373–387.

Heydari, A., Tavakoli, M., & Salim, N. (2016). Detection of fake opinions using time series. *Expert Systems with Applications, 58*, 83–92.

Horrigan, J. B., & Vitak, J. (2008). *The Internet and consumer choice: online Americans use different search and purchase strategies for different goods.* Pew Internet & American Life Project.

Jindal, N., & Liu, B. (Liu, 2007a). *Analyzing and detecting review spam. Data mining, 2007. ICDM 2007. seventh IEEE international conference on.* IEEE547–552.

Jindal, N., & Liu, B. (Liu, 2007b). *Review spam detection. Proceedings of the 16th international conference on world wide web.* ACM1189–1190.

Jindal, N., & Liu, B. (2008). *Opinion spam and analysis. Proceedings of the 2008 international conference on web search and data mining.* ACM219–230.

Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2016). Bag of tricks for efficient text classification. *CoRR* abs/1607.01759

Kouzis-Loukas, D. (2016). *Learning scrapy.* Packt Publishing Ltd.

Lai, M., Tambuscio, M., Patti, V., Ruffo, G., & Rosso, P. (2017). *Extracting graph topological information and users' opinion. 10456*, Springer112–118).

Li, F., Huang, M., Yang, Y., & Zhu, X. (2011). *Learning to identify review spam. IJCAI proceedings-international joint conference on artificial intelligence3. IJCAI proceedings-international joint conference on artificial intelligence* 2488.

Li, H., Chen, Z., Mukherjee, A., Liu, B., & Shao, J. (2015). *Analyzing and detecting opinion spam on a large-scale dataset via temporal and spatial patterns. ICWSM*634–637.

Li, Y., Feng, X., & Zhang, S. (2016). *Detecting fake reviews utilizing semantic and emotion model. Information science and control engineering (ICISCE), 2016 3rd international conference on.* IEEE317–320.

Liu, P., Xu, Z., Ai, J., & Wang, F. (2017). *Identifying indicators of fake reviews based on spammer's behavior features. Software quality, reliability and security companion (QRS-C), 2017 ieee international conference on.* IEEE396–403.

Luca, M., & Zervas, G. (2016). Fake it till you make it: Reputation, competition, and yelp review fraud. *Management Science, 62*(12), 3412–3427.

Mayzlin, D., Dover, Y., & Chevalier, J. (2014). Promotional reviews: An empirical investigation of online review manipulation. *American Economic Review, 104*(8), 2421–2455.

Mudambi, S. M., & Schuff, D. (2010). Research note: What makes a helpful online review? a study of customer reviews on amazon. com. *MIS Quarterly,* 185–200.

Mukherjee, A., Venkataraman, V., Liu, B., & Glance, N. S. (2013). *What Yelp fake review filter might be doing? Icwsm*409–418.

Ott, M., Choi, Y., Cardie, C., & Hancock, J. T. (2011). *Finding deceptive opinion spam by any stretch of the imagination. Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1.* Association for Computational Linguistics309–319.

Ouyang, Y., Li, W., Li, S., & Lu, Q. (2011). Applying regression models to query-focused multi-document summarization. *Information Processing & Management, 47*(2), 227–237. https://doi.org/10.1016/j.ipm.2010.03.005.

Park, D.-H., & Kim, S. (2008). The effects of consumer knowledge on message processing of electronic word-of-mouth via online consumer reviews. *Electronic Commerce Research and Applications, 7*(4), 399–410.

Park, J., Gu, B., & Lee, H. (2012). The relationship between retailer-hosted and third-party hosted wom sources and their influence on retailer sales. *Electronic Commerce Research and Applications, 11*(3), 253–261.

Petz, G., Karpowicz, M., F"urschu, H., Auinger, A., Stříteský, V., & Holzinger, A. (2014). Computational approaches for mining user's opinions on the web 2.0. *Information Processing & Management, 50*(6), 899–908. https://doi.org/10.1016/j.ipm.2014.07.005.

Poria, S., Cambria, E., & Gelbukh, A. (2016). Aspect extraction for opinion mining with a deep convolutional neural network. *Knowledge-Based Systems, 108*, 42–49.

Potthast, M., & Becker, S. (2010). *Opinion summarization of web comments. European conference on information retrieval.* Springer668–669.

Riegner, C. (2007). Word of mouth on the web: The impact of web 2.0 on consumer purchase decisions. *Journal of Advertising Research, 47*(4), 436–447.

Sotiriadis, M. D., & Van Zyl, C. (2013). Electronic word-of-mouth and online reviews in tourism services: the use of twitter by tourists. *Electronic Commerce Research, 13*(1), 103–124.

Tan, C., Lee, L., Tang, J., Jiang, L., Zhou, M., & Li, P. (2011). *User-level sentiment analysis incorporating social networks. Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining.* ACM1397–1405.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of Language and Social Psychology, 29*(1), 24–54.

Vapnik, V., Golowich, S. E., & Smola, A. J. (1997). *Support vector method for function approximation, regression estimation and signal processing. Advances in neural information processing systems*281–287.

Wang, G., Xie, S., Liu, B., & Yu, P. S. (2011). *Review graph based online store review spammer detection. 2011 iEEE 11th international conference on data mining*1242–1247. https://doi.org/10.1109/ICDM.2011.124.

Wang, S., & Manning, C. D. (2012). *Baselines and bigrams: Simple, good sentiment and topic classification. Proceedings of the 50th annual meeting of the association for computational linguistics: Short papers - volume 2ACL '12*Stroudsburg, PA, USA: Association for Computational Linguistics90–94.

Weise, K. (2011). A lie detector test for online reviewers. *Bloomberg Business Week*.

Zhang, D., Zhou, L., Kehoe, J. L., & Kilic, I. Y. (2016). What online reviewer behaviors really matter? effects of verbal and nonverbal behaviors on detection of fake online reviews. *Journal of Management Information Systems, 33*(2), 456–481. https://doi.org/10.1080/07421222.2016.1205907.

Zhu, F., & Zhang, X. (2006). The influence of online consumer reviews on the demand for experience goods: The case of video games. *ICIS 2006 Proceedings,* 25.

Ziegler, C.-N., & Skubacz, M. (2006). *Towards automated reputation and brand monitoring on the web. Web intelligence, 2006. wi 2006. IEEE/WIC/ACM international conference on.* IEEE1066–1072.