STUDY OF PERSONAL REVIEWS OF MOBILE USERS

Abstract

Users are now able to provide feedback on a wide range of service providers at any time, thanks to the proliferation of mobile apps for internet-connected devices. Nonetheless, sadly, to date, only a small number of classification technique researches have been implemented in this field. In this article, we analyzed more than 1,400,000 evaluations of actual mobile apps and found the following distinguishing features: There is a significant polarity difference, a short average length, a large length span, a power-law distribution, and a power-law distribution. Several studies have compared various emotion categorization algorithms, feature representations, and review times based on the aforementioned criteria.

Keywords: service providers, mobile apps, short average length, large length span, power-law distribution.

Authors

Mr. M. Ramnath
Department of Artificial Intelligence and Data Science Ramco Institute of Technology Rajapalayam,
Tamil Nadu, India.
ramnath25@gmail.com

Dr. C. Yesubai Rubavathi
Department of Computer Science and EngineeringFrancis Xavier Engineering College, Tirunelveli
Tamil Nadu, India.
yesubairubavathi@francisxavier.ac.in

I. INTRODUCTION

According to a new analysis from Global Digital Forensics, mobile devices like smartphones andtablets will soon overtake desktop computers for both professional and personal usage. Improving our IQ and gaining useful knowledge via smarter big data analysis is a top priority. Internet distributed computing has progressed rapidly in recent years, allowing us to analyze vast amounts of data and make accurate predictions about consumer preferences and future needs.

Understanding customers' sentiments and preferences via their written feedback is becoming more crucial. Some researchers have approached the analysis of customers' emotional inclination as a polarity classification problem; for instance, Turney1 used part-of-speech tagging to determine users' emotional inclination in their comments about cars, and Pang2 discovered that applying a unigram model to movie comments yielded accurate polarity classifications. While Tong3 used historical graphs to track their good and negative remarks, Dave4 used a grading algorithm, while Liu5 worked on Amazon and Epinions comments.

These findings open the way for emotion identification systems that use cognitive machine learning.

The shift from PC to mobile phone as the final platform has been gradual, thanks to the fast growth of Smart Phones. Mobile phone commenting gives consumers greater freedom. By 2019, the expected 5.6 billion Smartphone users will generate a data stream equal to 10 Exabytes (1018bits). Therefore, it is critical to apply text mining to efficiently extract the relevant information from massive datasets.

However, it's unclear whether conventional techniques for identifying emotional states can be reliably applied to evaluations of mobile apps. We are attempting to examine this via a series of comparison tests, since it has been mentioned in just a small number of literatures.

Here, we provide a brief overview of the paper's most important findings: Reviews made through mobile devices were the focus of the statistical study. Firstly, typical smartphone evaluations are far shorter than PC ones, clocking in at only 17 Chinese characters. Third, the length and number of reviews follow a power-law distribution, and fourth, there are three distinct polarities (positive, negative, and neutral). The reviews cover a wide range in length, from a single character to as many as 6,000. These statistical characteristics are crucial for developing the experimental procedure and implementing the adaptive categorization.

Better approaches for brief text comment have been sought via a number of comparative trials. Among these trials is a comparison of two different polarity classification algorithms (the tried- and-true SVM and the Naive Bayes approach). We compared the results of using the N-grammar model6,7,8 with N ranging from 1 to 4; (3) the influence of different count of word of comments on above experiments by splitting the data into groups based on word count to see how short textor long text influenced the results. These tests are useful in our search for a more precise and time-saving approach to emotion categorization.

The data used in the aforementioned comparison comes from a large-scale, real-world dataset. The experiment result has proven genuine reference relevance since it is based on a text corpus consisting of 140,000+ real mobile reviews scraped from iTunes for sentiment analysis. This scale level is beyond most of the text corpuses in current short text literature.

Merging, applying, and securing multimedia large data for various terminals is also crucial. To better track consumer behavior, gather product opinions, and safeguard against malicious calumniations and mischievous gossip spread by hostile rivals, this study is adapting the typical PC data processing onto mobile apps with certain appropriate alterations.

II. RELATED WORKS

The fields of natural language processing and data mining have paid a lot of attention to sentiment polarity categorization in recent years, particularly in the following ways:

One school of thought employs a symbolic method that involves 9-11 applying hand-created rules and a language. Sentiment lexicons are the backbone of lexicon-based approaches, which are excellent at knowledge-based categorization but suffer from a lack of generalizability due to their focus on specific domains. The polarity of emotion, on the other hand, may be seen as a text classification-based subject, making it a broad answer that is unrelated to any one area of study. Any text classification technique is then applied to the feature vectors representing the reviews. All the procedures described above fall within the second category.

Pang and Lee's2 research use data from IMDB user reviews (http://www.cs.cornell.edu/people/pabo/movie-review- data/), with each review averaging 3,500 bytes (or around 700 words) in length.

Using customer feedback evaluations (textual online discussion forums), Fan et al.12 provide a decision support for the finding of car defects. The study text corpus is drawn from the automobile safety complaint database maintained by the Office of Defect Investigations of the National Highway Traffic Safety Administration (NHTSA) of the United States Department of Transportation in 2010. Minimum required review length is 50 words, maximum review length is 8586 words, and an average review length is 502.

Electronic product reviews, such as those for digital cameras, computers, PDAs, MP3 players, etc., are compiled by Cui and Mittal13 from Froogle. The entire size of the corpus is close to 0.4GB, and it contains around 320k reviews of more than 80k distinct goods. The typical review length is 875 bytes long.

In his study14, Kasthuriarachchy analyzed the reviews he collected from many sources (movies, DVDs, phones, and tweets) to identify any discrepancies in the way they were classified semantically. The average amount of words in a sentence in the dataset is 17.2, with at least one phrase per review; the average number of sentences in a review of a film is 35.8.

From the above, it is clear that mobile app evaluations are much shorter than academic studies. (Please refer to subsection C for more information). Brief evaluations like this from the early stages of a research are often overlooked or deleted because they lack

sufficient context or read too much like spam. However, about 80% of all evaluations are of mobile applications, and they are often much shorter. When it comes to processing brief texts, how useful are conventional approaches and algorithms? It is important to compare and contrast existing methods and algorithms, thus it is important to have a discussion about the aspects of mobile app reviews and conduct a series of tests.

III. STATISTICS OF MOBILEUSER REVIEWS

1. **Dataset Collection:** We scoured the Apple App Store for evaluations left by WeChat's mobile users between 2021- 01-21 and 2023-03-06**.**

    - **Polarity allocation of feedback from customers:** iTunes reviews may be rated on a scale from 1 to 5. Reviews with a score of 4 or 5 are grouped together as the positive review class, reviews with a score of 1 or 2 are grouped together as the negative review class, and reviews with a score of 3 are grouped together as the neutral review class. The reviews are broken down by category in Table I. There is a total of 145,263 experimental data points, with 75.67 percent coming from positive examples,

    - **16.28 percent fromnegative examples, and 8.05 percent from neutral data.**

Table 1: The ratio of favorable to negative to neutral feedback

	Positive reviews	**Negativereviews**	**Neutralreviews**
count	109919	23656	11688
percentage	75.67%	16.28%	8.05%

2. **Evaluations' Statistical Characteristics:** Based on our statistical analysis of mobile user evaluations, we found:

    - **Much less word count compared to PC:** reviews Based on our experimental data, the average length of a review is 17 Chinese characters, much less than the average length of a Microblog post (45 words).

    - **Extremely extended duration compared to PC :**The smallest review in our data set consists of a single Chinese character, while the largest has almost 6,000. When it comes to text length, the Power law distribution holds true.

      The traditional power-law distribution holds true here, with most reviews consisting of a small number of words and a small number of reviews consisting of a big number of words. Power laws are shown in Figure 1(a) as relations of the form y kx, where y is the total number of reviews and x represents the review's individual character.The power law exponent is found to be -4.71 using experimental data.

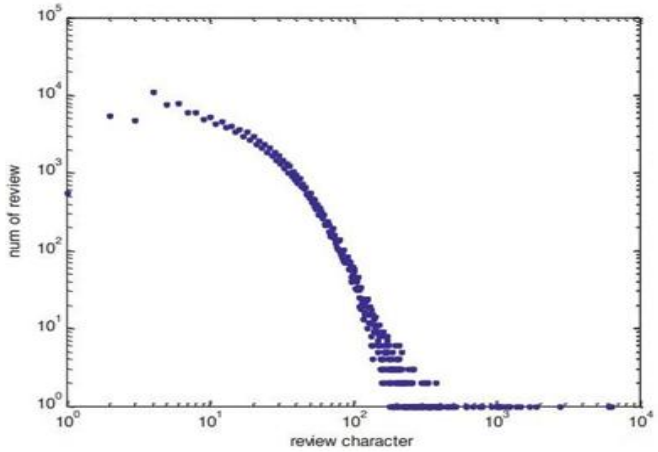

Figure 1 (a): The correlation between review count and review duration follows a power law distribution.

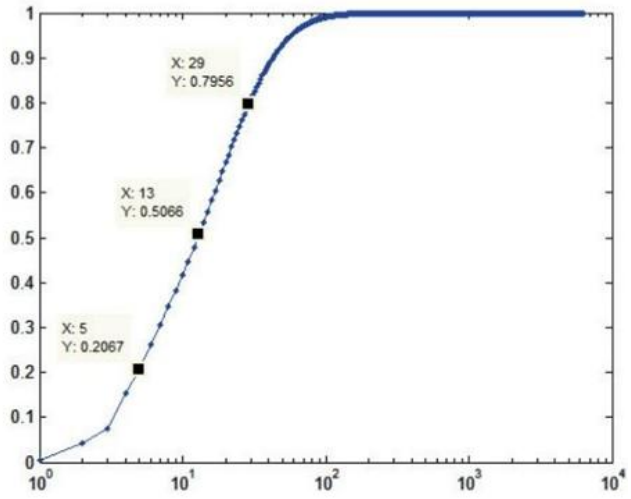

Figure1 (b): Review duration and review count cumulative distribution functions

Fig 1 (a) The correlation between review count and review duration follows a power law distribution.

The distribution of total documents is seen in (b) of Figure 1. Twenty percent of all reviews had less than five Chinese characters, fifty percent contained fewer than thirteen Chinese characters, and eighty percent contained fewer than twenty-nine Chinese characters, according to this data.

Thus, it is shown that there is a huge disparity in the lengths of reviews given and received, with the former receiving the vast bulk of attention.

Overall text length, word count, and no-repeat word count all vary significantly across favorable, negative, and neutral evaluations. Table 2 has more detailed statistical information.

Table 2: Experiment results shown as a frequency table with measures of overall length, number of words, and number of features

	DocLen (Total bytes count of review)			**DocWordCount (word count in review)**			**DocFeatureCount (no-repeat word count in review)**		
	Min/Max	***Mean***	***Std***	***Min/Max***	***Mean***	***Std***	***Min/Max***	***Mean***	***Std***
Positive	1~1891	15.75	21.85047	1~966	13.37	12.87804	1~380	8.623	7.60466
Neutral	1~355	30.54	24.74218	1~218	21.91	14.73096	1~120	14.72	9.83856
Negative	1~6336	34.11	65.96021	1~4611	24.69	43.32424	1~428	16.06	11.048
Total	1~6336	19.98	34.39604	1~4611	15.898	21.665	1~428	10.32	8.975

IV. EXPERIMENTAL WORKS

The primary text categorization procedure is shown in Figure 2. Before classification methods are used to the raw review data, it will be transformed into a feature vector representation; if the document features are too many, text feature extraction (selection) may be employed instead. This article conducts a series of experiments to compare the three major sub-processes shown in Figure 3 and primarily analyzes the following issues:

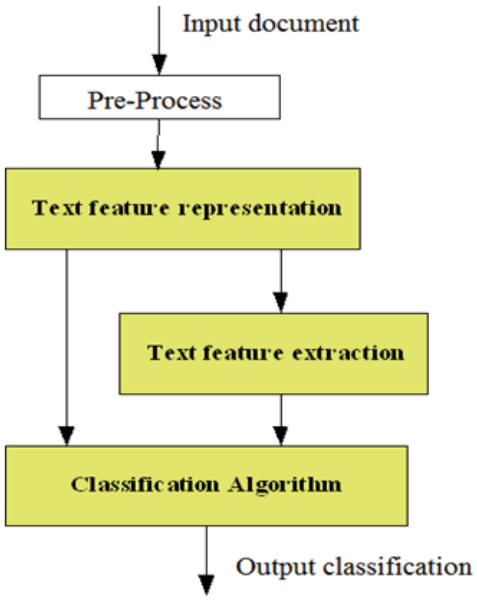

Figure 2: Architecture of Proposed work

To what extent can large collections of short texts be classified according to their polarity in terms of sentiment using traditional classification algorithms?

Researchers have shown that the Support Vector Machines technique and Nave Bayes4,15,16 work particularly well and consistently when used to sentiment categorization using machine learning. We'll be comparing these two tried-and-true approaches to smartphone reviews to see which works best for large collections of concise prose.

The formal approach of text feature representation for input data. Characters, words, phrases, and even whole sentences are all valid data types. While words serve as clear boundaries in English, they are sometimes ill-defined in Chinese. The material is segmented via the Chinese word segmentation method, which always uses phrases or words to represent the text. However, mobile app evaluations are rife with internet lingo, mistakes, and colloquial vocabulary, makingit challenging for the standard Chinese Word Segmentation algorithm to appropriately split the words or phrases. In addition, the present methods of Chinese word segmentation focus more on separating words and phrases according to their parts of speech than on separating them according to their meaning. Even though "very good" is broken down into "very" and "good" in English, that's really a compliment in Chinese. In addition, the negative connotation of "not very well" is reflected in the Chinese dissection of the phrase into its component parts. As a consequence, the Chinese Word Segmentation method may have a difficult time telling between two polar opposite sentences since their words may be represented similarly.

To address this issue, the N-gram language model may be used to streamline the Chinese Word Segmentation and close this void. However, the recall rate for text categorization drops as N in N-gram increases, since the feature vector of the document becomes redundant. Therefore, we need to be more selective while choosing N. Here, we want to experiment with and validate the enormous short text corpus to identify the optimal fit value of N. Will the aforementioned techniques continue to be useful for a range of review times?

There is a huge variety in review length, with some being 6,000 characters or more and others having only one. Will the aforementioned techniques continue to be useful for a range of review times? To address this issue, we developed two testing procedures:

The size of a review is used to categorize it. We divided the evaluations up into ten categories: 1+, 10+, 20+, 30+, 50+, 70+, 100+, 150+, 200+, and 300+. Reviews having more than one Chinese word are denoted by 1+, while reviews with more than 300 Chinese words are denoted by 300+.

The other is organized by sample size. To avoid sample size bias in the categorization, we randomly split the reviews into four groups of equal size.

In order to provide a thorough comparison of the respective performances, this research chooses two assessment indices. One such metric is the F- SCORE, a score that balances recall with accuracy. Area under the receiver operating characteristic curve is another metric used for assessment. In the experiment; the 10-fold cross-validationtechnique is used.

V. RESULTS AND DISCUSSIONS:

In this work, the SVM (LibLinear) and Naive Bayes (Multinomial) methods are contrasted. The Confusion Matrix for this evaluation is shown in Table3. Comparison of the two approaches is shown in Table 4. Based on the P-R-F index, we know that Naive Bayes Multinomial is superior at negative and neutral classification whereas LibLinear performs better at positive classification. While both approaches are very reliable, the AUC index favors Nave Bayes Multinomial classification over liblinear. Neither approach is very effective in finding neutralclassifications, and this is directly connected to the size of the text corpus.

Table 3: SVM LibLinear and Naive Bayes Multinomial Confusion Matrix

	Negative	SVM Lib Linear Neutral	Positive	Negative	Naïve Bayes Multinomial Neutral	Positive
Negative	15809	600	7245	16994	4097	2563
Neutral	3531	644	7513	3818	4259	3611
Positive	3330	784	105787	4488	8766	96647

Table 4: The SVM LibLinear and Naive Bayes Multinomial Classification Methods: A Comparison ofClassification Results

polarity	SVM		LibLinear		Naïve Bayes Multinomial			
	P	*R*	*F*	*AUC (ROC area)*	*P*	*RF*	*AUC(ROC area)*	
Positive	0.878	0.963	0.918	0.772	0.94	0.879	0.909	0.922
Negative	0.697	0.668	0.683	0.806	0.672	0.718	0.694	0.934
Neutral	0.318	0.055	0.094	0.522	0.249	0.364	0.296	0.81
Total	0.803	0.842	0.813	0.758	0.841	0.812	0.824	0.915

Grouping by review word count allows for a comparison of two approaches, which is shown in Table 5. Naive Bayes Multinomial is superior than SVM LibLinear because its general index is greater for word counts of 200 or less.

Table 5: Group comparison of two approaches based on total review duration

Reviews word count	**Precision**	**SVM LibLinear Recall**	**F-Measure**	**Precision**	**Naïve Bayes Multinomial Recall**	**F-Measure**
1+	0.803	0.842	0.813	0.841	0.812	**0.824**
10+	0.769	0.812	0.780	0.812	0.780	**0.794**
20+	0.693	0.740	0.700	0.732	0.703	**0.715**
30+	0.649	0.699	0.658	0.687	0.663	**0.673**
50+	0.610	0.660	0.624	0.656	0.635	**0.644**

70+	0.582	0.636	0.600	0.635	0.617	**0.625**
100+	0.584	0.626	0.594	0.601	0.597	**0.598**
150+	0.505	0.545	0.512	0.589	0.597	**0.568**
200+	0.439	0.473	0.443	0.519	0.550	**0.510**
300+	0.559	0.583	**0.569**	0.542	0.563	0.488

See Table 6 for a comparison of the two approaches based on our N-gram model experiments with N values between 1 and 4.

Table 6: Sample sizes (N) for each of the two approaches were anything from one to four.

N-gram mar model	**Precision**	**SVM LibLinear Recall**	**F-Measure**	**Naïve Bayes Multinomial Precision**	**Recall**	**F-Measure**
n=1 gram	0.803	0.842	0.813	0.841	0.812	0.824
n=2 gram	0.813	0.846	**0.824**	0.854	0.810	**0.828**
n=3 gram	0.812	0.845	0.823	0.852	0.808	0.827
n=4 gram	0.810	0.844	0.812	0.847	0.808	0.820

Table6 shows that for N=2, both methods have generally acceptable classification effects. The categorization impact is not enhanced but rather diminished when N is increased to 3 or 4 grams. Therefore, the optimal choice is 2-gram. Then, we divided the reviews into two groups based on the total word count of each. The results of a direct segmentation of Chinese words are shown in Table 7. Sentiment polarity classification accuracy increases with decreasing word count for both approaches, as seen by the results. Meanwhile, the 2-grammer shows the most improvement in the categorization task. Classification precision is directly proportional to review length, with longer reviews being more challenging to categorize.

Table 7: The number of reviews (N) used to compare the two approaches varied from one tofour.

Revie ws word count	LibLinear F-Measure	1gram Multinomial F-Measure	2gram LibLinear Multinomial F-Measure	F-Measure	LibLinear F-Measure	3gram Multinomial F-Measure	4gram LibLinear Multinomial F-Measure	F-Measure
1+	0.813	0.824	0.824	**0.828**	0.824	0.826	0.824	0.825
10+	0.78	0.794	0.791	**0.797**	0.791	0.796	0.79	0.793

20+	0.7	0.715	0.71	**0.722**	0.709	0.721	0.708	0.716
30+	0.658	0.673	0.667	**0.683**	0.666	0.678	0.665	0.667
50+	0.624	0.644	0.629	**0.647**	0.629	0.637	0.627	0.629
70+	0.6	0.625	0.613	**0.629**	0.612	0.608	0.607	0.602
100+	0.574	0.578	0.595	**0.593**	0.576	0.577	0.578	0.575
150+	0.512	**0.568**	0.491	0.56	0.491	0.56	0.491	0.561
200+	0.443	0.51	0.506	**0.512**	0.455	0.507	0.457	0.493
300+	**0.569**	0.488	0.546	0.533	0.527	0.464	0.505	0.449

To assess how sample size impacts performance, we informally categorize the experimental information at hand into four groups of varying durations. There are a total of 38198 reviews with 1–6 characters selected by the top group. The second column of Table 8 displays the number of reviews allotted to each of the remaining three categories.

Table 8 displays the results of our 10-fold cross-validation procedure, which we applied to four different groups before conducting an experimental comparison of the two primary algorithms using 1 gram and 2 gram language models. These data show that when the total number of words in reviews increases, the F-measures for both approaches drop. And F-measure has the best outcome when N=2. Sentiment analysis performed on Chinese literature shows that the information included in shorter passages is both highly focused and very effective. The accuracy and recall ratios may continue to drop if the text is extensive and contains numerous non- sentiment terms that interfere with the detection of sentiment polarity.

Table 8: When the sample size, N, ranges from 1 to 4, we compare the two approaches in eachgroup

Group	Reviews count	1gram LibLinearMultinomial F-Measure	F-Measure	2gram LibLinearMultinomial F-Measure	F-Measure	LibLinear F-Measure	3gram Multinomial F-Measure	LibLinear F-Measure	4gramMultinomial F-Measure
1#(1-63	*38198*	0.943	0.948	0.95	**0.967**	0.931	0.944	0.915	0.926
2#(7-133	*36999*	0.88	0.885	0.886	**0.898**	0.85	**0.898**	0.85	0.87
3#(14-253	*35214*	0.756	0.771	0.761	**0.774**	0.761	0.774	0.754	0.761
4#(25+3	*34854*	0.65	0.673	0.654	**0.687**	0.654	0.687	0.65	0.672

VI. CONCLUSION

We crawled a massive amount of actual data, ran statistical analysis, and discovered that evaluations of mobile applications have four common traits that make them a good source of information: One has a power-law distribution with a short average length and a long span of length. There is no discernible flip in polarity. Due to these distinctions, evaluations on mobile devices stand out from their PC counterparts.

This research compares the frequently used classical algorithm and its processing with a variety of experimental setups to determine which semantic categorization approaches are best suited formobile reviews. The experiments show that: (1) when the number of words in reviews is greater than 150, the feature-extracting process must occur and can obviously improve sentiment categorization accuracy; (2) when the Chinese word segmentation is complete, the best result is achieved by using N-Gram (N=2) for feature representation; (3) when the number of words in reviews is greater than 150, the fewer words there are to classify, the better the results; (4) we found that the fewer words there are to classify, the better the results. Reviews get increasingly challenging to categorize in proportion to the number of words they include. Based on these features, as described in this research, and our past experimental findings, we want to explore a complete approach tailored specifically for mobile application review classifications.

REFERENCES

[1] Abilhoa WD, De Castro LN (2014) A keyword extraction method from twitter messagesrepresented as graphs. Appl Math Comput 240:308–325
[2] Ahmad M, Aftab S, Bashir MS, Hameed N (2018) Sentiment analysis using SVM: asystematic literature review. Int J Adv Comput Sci Appl 9(2)
[3] Alantari HJ, Currim IS, Deng Y, Singh S (2022) An empirical comparison of machine learning methods for text-based sentiment analysis of online consumer reviews. Int J Res Mark 39(1):1–19
[4] Alessia D, Ferri F, Grifoni P, Guzzo T (2015) Approaches, tools and applications for sentiment analysis implementation. Int J Comput Appl 125(3):26–33
[5] Alfter D, Cardon R, François T (2022) A dictionary-based study of word sense difculty. In: Proceedings of the 2nd workshop on tools and resources to empower people with REAding DIfculties (READI) within the 13th language resources and evaluation conference. European Language Resources Association, pp 17–24
[6] Altheneyan AS, Menai MEB (2014) Naïve bayes classifers for authorship attribution of Arabic texts. J King Saud Univ-Comput Inf Sci 26(4):473–484
[7] Antypas D, Preece A, Collados JC (2022) Politics and virality in the time of twitter: a large-scale crossparty sentiment analysis inGreece, Spain and united kingdom. arXiv preprint arXiv:2202.00396
[8] Appel O, Chiclana F, Carter J, Fujita H (2016) A hybrid approach to the sentiment analysis problem at the sentence level. Knowl-Based Syst 108:110–124
[9] Athanasiou V, Maragoudakis M (2017) A novel, gradient boosting framework for sentiment analysis in languages where NLP resources are not plentiful: a case study for modern Greek. Algorithms 10(1):34
[10] Baccianella S, Esuli A, Sebastiani F (2010) Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In: Proceedings of the seventh international conference on language resources and evaluation (LREC'10), vol 10. European Language Resources Association (ELRA), pp 2200–2204
[11] Bahrainian S-A, Dengel A (2013) Sentiment analysis and summarization of twitter data. In: 2013 IEEE 16th international conference on computational science and engineering. IEEE, pp 227–234
[12] Baid P, Gupta A, Chaplot N (2017) Sentiment analysis of movie reviews using machine learning techniques. Int J Comput Appl 179(7):45–49
[13] Balahur A, Hermida JM, Montoyo A (2011) Building and exploiting emotinet, a knowledge base for emotion detection based on the appraisal theory model. IEEE Trans Afect Comput 3(1):88–101

[14] Banea C, Mihalcea R, Wiebe J (2014) Sense-level subjectivity in a multilingual setting. Comput Speech Lang 28(1):7–19
[15] Bao H, Li Q, Liao SS, Song S, Gao H (2013) A new temporal and social PMF- based method to predict users' interests in micro- blogging. Decis Support Syst 55(3):698–709.
[16] Statcounter (2020), "Desktop vs Mobile vs Tablet Market Share Worldwide," (accessed October 15, 2021), https://gs.statcounter.com/platform-market-share/desktop-mobile-tablet.
[17] SurveyMonkey (2020), "Mobile Optimization for Surveys: A How-to Guide," (accessed October 15, 2021), https://www.surveymonkey.com/mp/mobile-optimization-for-surveys/.
[18] Sela Aner, Simonson Itamar (2017), "The Feeling of Preference," working paper, https://ssrn.com/abstract=3384177.
[19] Sela Aner, Hadar Liat, Morgan Siân, Maimaran Michal (2019), "Variety-Seeking and Perceived Expertise," Journal of Consumer Psychology, 29 (4), 671–79.
[20] PwC (2020), "Global Consumer Insights Survey 2020," research report, https://www.pwc.com/gx/en/consumer-markets/consumer-insights-survey/2020/pwc-consumer-insights-survey-2020.pdf.

