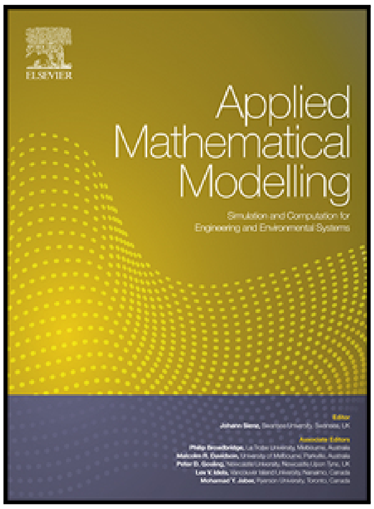
Accepted Manuscript



ReviewModus: Text Classification and Sentiment Prediction of Unstructured Reviews using a Hybrid Combination of Machine Learning and Evaluation Models

Fouad Zablith , Ibrahim H. Osman

PII: S0307-904X(19)30117-9
DOI: https://doi.org/10.1016/j.apm.2019.02.032
Reference: APM 12685

To appear in: *Applied Mathematical Modelling*

Received date: 11 July 2018
Revised date: 14 December 2018
Accepted date: 25 February 2019

Please cite this article as: Fouad Zablith , Ibrahim H. Osman , ReviewModus: Text Classification and Sentiment Prediction of Unstructured Reviews using a Hybrid Combination of Machine Learning and Evaluation Models, *Applied Mathematical Modelling* (2019), doi: https://doi.org/10.1016/j.apm.2019.02.032

Highlights

- We propose a novel predictive analytics framework to classify and analyze unstructured reviews
- Evaluation models can potentially train machine learning algorithms for predicting reviews classification and sentiments
- We implement neural networks and logistic regression algorithms tested in the context of e-government service evaluation
- The classification reached a promising F-score of 85.16%, and sentiments correlating 71.44% with a manually validated dataset
- The framework contributes to uncover hidden insights that were not initially captured by closed-ended questionnaires

ReviewModus: Text Classification and Sentiment Prediction of Unstructured Reviews using a Hybrid Combination of Machine Learning and Evaluation Models

Fouad Zablith*, Ibrahim H. Osman

Olayan School of Business,
American University of Beirut
PO Box 11-0236,
Riad El Solh, 1107 2020,
Beirut, Lebanon

* Corresponding author.
E-mail addresses: fouad.zablith@aub.edu.lb (F. Zablith), ibrahim.osman@aub.edu.lb (I. H. Osman).

Abstract

While research interest on product and service evaluation from unstructured text reviews is increasing, investigating the effectiveness of predictive analytical models in this context is still underexplored. With the advancement in machine learning research, an opportunity exists to bridge this gap using a model-based product and service evaluation. We propose in this article *ReviewModus*, a text mining and processing framework that (1) relies on the model structure and its corresponding assessment questions to train a machine learning algorithm to predict the classification of reviews around the model dimensions; (2) predicts the sentiments within the reviews based on external review training datasets; and (3) transforms the extracted measures from the reviews for further analysis. Our approach is evaluated in the context of 11 e-Government services where the performance of the framework is compared to the manual processing of unstructured reviews cross-checked by three independent evaluators. Our study shows promising classification results with a micro-average F-score reaching 85.16%, and a high sentiment prediction correlation (71.44%) with the manually performed sentiment assessment.

Keywords: machine learning; text mining; neural network; logistic regression; e-government

1. Introduction

Traditional evaluation and user satisfaction models have been extensively used to analyze usersô feedback provided in structured forms [1ï3], however the surge of usersô opinions in unstructured forms is still a challenge to such models. Researchers have invested a good amount of effort in developing and testing evaluation models to translate the data collected from usersô feedback into meaningful actions. Such models typically take a set of dimensions relevant to the product or service being assessed as input, and feed them into appropriate output dimensions to generate evaluation measures. The input and output variables are usually defined around the product or service characteristics and are often populated by designing a set of survey questions. Such models play a major role in testing certain hypotheses and ensuring a consistent evaluation process across different products and services. However, they usually require a pre-defined and structured data input which is feasible in controlled data collection settings, achieved through closed-ended questions [1], but is much harder to realize through text-based reviews.

Compared to the traditional challenges associated with prompting users to fill questionnaires [4], todayôs internet savvy users freely give away their opinions online and through social media platforms, mainly in an unstructured way [5]. Such online platforms are becoming the de-facto channels for reporting usersô concerns and product acceptance at an ñunprecedented scale in real-timeò [6]. As a result, traditional well-defined product and service evaluation processes require more accommodation of the real-time and dynamic aspect of todayôs opinion sharing channels. While we are witnessing an increased research interest in opinion mining from text [7ï10], most of the available approaches do not incorporate the existing structure of well-established product evaluation and user satisfaction models in their methodologies. Our research aims to close this gap, with a focus on the following research question: *how can we accommodate product and service evaluation models in the process of automatically analyzing unstructured users' reviews?*

To answer this research question, we propose in this article ñ*ReviewModus*ò, a model-based supervised machine learning framework, to assist with the automatic extraction and analysis of measurable variables from unstructured text in product and service reviews. The novelty of our approach stems from the augmentation of the evaluation process of unstructured usersô reviews by using traditional questionnaire evaluation methods as a means for training a predictive machine learning algorithm. The framework learns to predict the classification of reviews into a pre-defined set of evaluation model dimensions, and to predict the degree of sentiments expressed in the reviews. The predicted classifications and sentiment measures are then processed for further analysis and generating actionable insights. The feasibility of our approach is demonstrated by investigating the use of a neural network-based (NN) algorithm. The NN algorithm is trained on a set of existing survey questions for understanding the pre-defined dimensions of a user satisfaction model. It is coupled with a logistic regression algorithm trained on Amazon.com reviews for predicting and quantifying sentiments expressed in the reviews. The evaluation is conducted in the domain of e-government service assessment. In this context, we employ a user satisfaction model that evaluates e-government services around the Cost, Benefit, Risk and Opportunity dimensions (i.e., COBRA model) [2]. Our evaluation shows promising results. Our tests demonstrate a promising micro average F-score of 85.16% with respect to the multiclass prediction of model dimensions, and a high positive correlation of 71.44% with the assessment of sentiments performed manually by three evaluators.

The remaining parts of the article are structured as follows. Section 2 provides a review of related works in the field. Section 3 presents our model-based supervised machine learning framework. Section 4 focuses on the evaluation procedure in the domain of e-Government services. Section 5 presents our results, and Section 6 concludes with future research directions.

2. Review of User Opinion Analysis: Model and Machine Learning Perspectives

In this section we review existing machine learning and model-based approaches to assess usersô opinions on products and services.

Turning information in text into ñactionable knowledgeò is increasingly getting research attention [11]. This attention is gaining momentum in various domains. For example, Reddick et. al. [12] investigate the impact of the analysis of text in social media on the delivery of public services; while Müller et. al. [13] study how text analytics can help in better understanding customersô problems and requests for improving customer service. The efforts involved in text processing are pushing for the automation of tasks related to text analysis. Such tasks range from sentiment detection in user generated content on the web [8] and question answering [14,15], to classification and prediction [11,16], to name a few. In this context, machine learning is being more and more involved in performing text analysis functionalities. For example, neural networks were used for classifying text documents [17]; Support Vector Machines (SVM) were heavily employed in pattern recognition [18], in processing customer reviews and product opinions [19], and in feature-based text categorization [20]; statistical and evolutionary algorithms were tested for Part-of-Speech tagging [21]; and a Bayesian approach was used to model customer satisfaction from unstructured text [7]. While those approaches are proving to be effective, some of the challenges remain pertinent to the success or failure of those techniques, including the type of data in focus and required preparation, selecting the right classification approach, and the availability of appropriate training data used for the algorithm [22]. We focus on the challenge involved in providing the supervision and training needed for the success of supervised machine learning algorithms in text-based review analytics.

Parallel to the efforts invested in machine learning and feature-based analysis, another flourishing area of research is studying models to represent and capture various analytical contexts and objectives. For example, researchers have been working on designing models to represent user satisfaction in the context of public services [1,3] and e-Government [2,23,24], while others have focused on modeling usefulness of technology-driven solutions in more generic terms [25,26]. The aim of such related works is to come up with well-defined models to represent the situation being assessed as accurately as possible, for testing certain hypotheses. The developed evaluation models usually involve a set of inter-related dimensions consisting of a set of *inputs*, which are transformed into a set of *outputs*. For instance, Parasuraman et. al. [1] highlighted the importance of *reliability* and *responsiveness* among other input dimensions to measure the service *quality* output dimension in their SERVQUAL model. For such models to perform well, analysts usually assess the importance of the model dimensions by controlling the collection of data around the dimensions to be evaluated. In the context of usersô feedback, such dimensions are often controlled through a set of closed-ended questions to ensure a consistent and measurable assessment of usersô perception of the dimensions in focus [27]. To ensure a holistic view of the participantsô feedback, service and evaluation processes often provide an option for participants to express their opinions using an open-ended format, which are subsequently analyzed. This option is provided for various reasons including serving as a ñsafety netò in support for the closed questions of the survey, or seeking further information from the participants on uncovered aspects in the other structured parts of the survey questions [28]. However, the complexity involved in analyzing the open-ended feedback often results in having a substantial amount of untapped text data, which could provide additional insights for service and product improvement.

Hence the question is how could the consistency and robustness of existing models be employed to analyze unstructured usersô reviews? While machine learning approaches on mining opinions from text are flourishing, we have seen little efforts on modeling frameworks that incorporate predictive algorithms in model-based approaches to assess text-based opinions. Our aim is to close this gap in the literature.

3. A Framework for Model-Based Supervised Machine Learning

As discussed in Section 2, model-based evaluation approaches provide a solid methodology and well-tested hypothesis for evaluating products and services around key dimensions and variables. We see an opportunity to exploit such model characteristics for automating unstructured review analysis. Most of such models are proposed and assessed based on questionnaires that involve meticulously crafted closed-ended questions, coupled with Likert scale format answers [27]. For example in SERVQUAL, one of the first models proposed to evaluate services, five dimensions were considered in service evaluation, namely: *Tangibles*, *Reliability*, *Responsiveness*, *Assurance* and *Empathy* [1]. Such dimensions were then tested using a set of Likert scale questions. For instance, *Reliability* was tested based on five questions such as ñwhen these firms promise to do something by a certain time, they should do so.ò Figure 1 provides an example of a connection between userôs input and a modelôs dimension through a survey question.

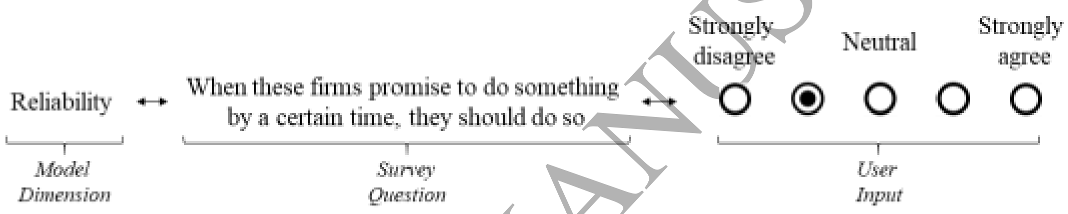


Figure 1 - Example of Collecting User Input on a SERVQUAL Model Dimension through a Closed-Ended Question

In the case of unstructured reviews, we assume that the presence of a model potentially gives an indication of how to interpret the reviews. In other words, the analyst can rely on the model to identify key elements to focus on while processing and coding reviews. The model provides the required semantics and structural mechanisms used for analyzing the service and product in focus. Semantics involve understanding the meaning and classification of content generated from the reviews; while the structural mechanisms involve the dynamics between the content elements. Irrespective of the types of supporting tools, we identify the need for a 3-phase framework to extract meaningful insights from text feedback around a pre-defined model:

- Phase 1: Classify text statements around the product or service variables that are represented in the pre-defined evaluation model.

- Phase 2: Identify the level of agreement and sentiment associated with the mentioned text statement.
- Phase 3: Extract quantifiable entities around text statements for analyzing, interpreting and mining insights.

In this context, to successfully perform Phase 1, the content analyst needs to have a clear understanding of the meaning of the evaluation model variables to be able to consistently classify the review statements. For example, a protocol with an explicit vocabulary for coding and classifying the statements can be developed for the analysts to follow. Concerning Phase 2, a clear methodology for sensing the level of agreement in a review statement is needed. For instance, a dictionary of keywords can be developed, or patterns in the text that indicate a support or disagreement with the stated text. With respect to Phase 3, the analyzed content must be translated through quantifiable measures including for example specified metrics around the evaluation model dimensions, such as the overall agreement level or other measures pertinent to the analytical goals. With the increase in the amount of text to analyze, performing these phases manually is a tedious and challenging task.

We propose in this article a ñReviewModusò framework, which combines model-based assessment and machine learning techniques to achieve new insights that cannot be obtained by either approach separately. The illustrated details of the framework are depicted in Figure 2. At a high level, the framework supports Phase 1 through the Classification Prediction component. Phase 2 is supported by the Sentiment Prediction component, and Phase 3 is realized by the Analytics component. The components are formed of three steps each.

the gap between the loose nature of unstructured text review analysis around the well-tested product and service evaluation models.

Our approach can benefit from further improvements at different levels. First, while the performance of the classification algorithm was promising, such results might fluctuate. This is largely due to the initial randomization of weights applied on the synapses of the neural network, coupled with the small training dataset used in our scenario. Our work can be extended to test and compare the performance of different machine learning techniques including for example Support Vector Machine [30], Ensemble Learning [31], and further neural network-based configurations such as deeper neural networks, Long Short Term Memory (LSTM) [46] or Character Level Convolutional Networks [47], and the possibility of using transfer learning [48] from different product evaluation and user satisfaction model scenarios. Second, while this work was tested in the context of e-Government services, the framework can benefit from further tests to perform in other domains. One interesting aspect is to study how the product and service context might impact the performance of the review analysis tasks. For example, in some contexts the designed questionnaires and related model might be high level, compared to reviews that capture more granular feedback, making the classification task more challenging. Third, one challenge related to the use of machine learning is the inability to investigate how results have been generated. A potential extension to our work is to complement this machine learning based approach with external background knowledge sources to provide a degree of reasoning behind the classification and sentiment analysis tasks. Fourth, further investigation is required at the level of improving the performance of our tested algorithms. As shown in the previous statement samples (e.g., Figure 9), the algorithm mis-classified some review statements. We anticipate that this can largely be due to the training phase of the algorithm used. One potential way to address this limitation is to investigate in the future the option of implementing a feedback loop during the training step of the algorithm, coupled with a monitoring process of the improvement of the fitness of the algorithm above the current acceptable correlation level of 71.44%.

In addition to further improving our approach, we are planning as part of our future work to check how the results extracted from the reviews correlate with the quantitative analysis performed through the survey questions around the COBRA dimensions. This presents a good research opportunity, given our access to both structured and unstructured data emanating from the same users assessing the same services. It would be interesting to investigate how these two approaches complement each other, and ultimately see to what degree can unstructured review analytics lift the burden imposed by conducting traditional survey methods.

To conclude, we see our proposed approach as a mean for augmenting the analystôs ability to make sense of the increasingly available unstructured user feedback guided by product evaluation and satisfaction models. One of the major contributions of our work is that it can possibly help uncovering hidden insights that were not initially captured by closed-ended questionnaires and pre-designed models. Furthermore, our proposed approach can potentially help improving and refining models that are more aligned with usersô expectations from products and services.

Acknowledgements

We would like to thank the editors and anonymous reviewers for their suggestions and invaluable feedback for improving our article. This work was supported by NPRP grant # [NPRP 09 - 1023 - 5 ï 158] from the Qatar National Research Fund (a member of Qatar Foundation), and the University Research Board (URB) grant at the American University of Beirut.

References

[1] A. Parasuraman, V.A. Zeithaml, L.L. Berry, SERVQUAL: a multiple-item scale for measuring consumer perceptions of service quality, Journal of Retailing; Greenwich. 64 (1988) 12ï40.

[2] I.H. Osman, A.L. Anouze, Z. Irani, B. Al-Ayoubi, H. Lee, A. Balcē, T.D. Medeni, V. Weerakkody, COBRA framework to evaluate e-government services: A citizen-centric perspective, Government Information Quarterly. 31 (2014) 243ï256. doi:10.1016/j.giq.2013.10.009.

[3] Alessandro Ancarani, Towards quality e-service in the public sector: The evolution of web sites in the local public service sector, Managing Service Quality. 15 (2005) 6ï23. doi:10.1108/09604520510575236.

[4] S.D. Crawford, M.P. Couper, M.J. Lamias, Web Surveys: Perceptions of Burden, Social Science Computer Review. 19 (2001) 146ï162. doi:10.1177/089443930101900202.

[5] V. Dhar, Data science and prediction, Communications of the ACM. 56 (2013) 64ï73.

[6] W. Duan, B. Gu, A.B. Whinston, Do online reviews matter? ð An empirical investigation of panel data, Decision Support Systems. 45 (2008) 1007ï1016. doi:10.1016/j.dss.2008.04.001.

[7] M. Farhadloo, R.A. Patterson, E. Rolland, Modeling customer satisfaction from unstructured data using a Bayesian approach, Decision Support Systems. 90 (2016) 1ï11. doi:10.1016/j.dss.2016.06.010.

[8] B. Pang, L. Lee, Opinion Mining and Sentiment Analysis, Foundations and Trends in Information Retrieval. 2 (2008) 1ï135. doi:10.1561/1500000011.

[9] Atika Qazi, Ram Raj, Glenn Hardaker, Craig Standing, A systematic literature review on opinion types and sentiment analysis techniques: tasks and challenges, Internet Research. 27 (2017) 608ï630. doi:10.1108/IntR-04-2016-0086.

[10] S. Moghaddam, M. Ester, Opinion Digger: An Unsupervised Opinion Miner from Unstructured Product Reviews, in: Proceedings of the 19th ACM International Conference on Information and Knowledge Management, ACM, New York, NY, USA, 2010: pp. 1825ï1828. doi:10.1145/1871437.1871739.

[11] R. Gopal, J.R. Marsden, J. Vanthienen, Information mining ð Reflections on recent advancements and the road ahead in data, text, and media mining, Decision Support Systems. 51 (2011) 727ï731. doi:10.1016/j.dss.2011.01.008.

[12] C.G. Reddick, A.T. Chatfield, A. Ojo, A social media text analytics framework for double-loop learning for citizen-centric public services: A case study of a local government Facebook use, Government Information Quarterly. 34 (2017) 110ï125.

[13] O. Müller, S. Debortoli, I. Junglas, J. vom Brocke, Using Text Analytics to Derive Customer Service Management Benefits from Unstructured Data, MIS Quarterly Executive. 15 (2016) 243ï258.

[14] S.P. Conlon, B.J. Reithel, M.W. Aiken, A.I. Shirani, A natural language processing based group decision support system, Decision Support Systems. 12 (1994) 181ï188. doi:10.1016/0167-9236(94)90002-7.

[15] D. Ravichandran, E. Hovy, Learning surface text patterns for a question answering system, in: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, 2002: pp. 41ï47. http://dl.acm.org/citation.cfm?id=1073092 (accessed November 14, 2016).

[16] F. Sebastiani, Machine learning in automated text categorization, ACM Computing Surveys (CSUR). 34 (2002) 1ï47.

[17] B. Yu, Z. Xu, C. Li, Latent semantic analysis for text categorization using neural network, Knowledge-Based Systems. 21 (2008) 900ï904. doi:10.1016/j.knosys.2008.03.045.

[18] C.J. Burges, A tutorial on support vector machines for pattern recognition, Data Mining and Knowledge Discovery. 2 (1998) 121ï167.
[19] K.-W. Cheung, J.T. Kwok, M.H. Law, K.-C. Tsui, Mining customer product ratings for personalized marketing, Decision Support Systems. 35 (2003) 231ï243. doi:10.1016/S0167-9236(02)00108-2.
[20] T. Joachims, Text categorization with support vector machines: Learning with many relevant features, in: European Conference on Machine Learning (ECML), Springer Berlin/Heidelberg, 1998: pp. 137ï142. http://link.springer.com/chapter/10.1007/BFb0026683 (accessed November 8, 2016).
[21] R. Forsati, M. Shamsfard, Hybrid PoS-tagging: A cooperation of evolutionary and statistical approaches, Applied Mathematical Modelling. 38 (2014) 3193ï3211.
[22] D.M. Christopher, R. Prabhakar, S. Hinrich, Introduction to information retrieval, 2008.
[23] D. Gouscos, M. Kalikakis, M. Legal, S. Papadopoulou, A general model of performance and quality for one-stop e-Government service offerings, Government Information Quarterly. 24 (2007) 860ï885. doi:10.1016/j.giq.2006.07.016.
[24] S. Rotchanakitumnuai, Measuring e-government service value with the E-GOVSQUAL-RISK model, Business Process Management Journal. 14 (2008) 724ï737. doi:10.1108/14637150810903075.
[25] V. Venkatesh, M.G. Morris, G.B. Davis, F.D. Davis, User acceptance of information technology: Toward a unified view, MIS Quarterly. 27 (2003) 425ï478.
[26] F.D. Davis, Perceived usefulness, perceived ease of use, and user acceptance of information technology, MIS Quarterly. 13 (1989) 319ï340.
[27] Nitin Seth, S.G. Deshmukh, Prem Vrat, Service quality models: a review, Int J Qual & Reliability Mgmt. 22 (2005) 913ï949. doi:10.1108/02656710510625211.
[28] A. OôCathain, K.J. Thomas, Any other comments? Open questions on questionnairesïa bane or a bonus to research, BMC Med Res Methodol. 4 (2004) 25.
[29] P. Soucy, G.W. Mineau, A simple KNN algorithm for text categorization, in: Proceedings of IEEE International Conference on Data Mining, IEEE, 2001: pp. 647ï648.
[30] T. Joachims, Learning to classify text using support vector machines: Methods, theory and algorithms, Kluwer Academic Publishers Norwell, 2002.
[31] L. Shi, X. Ma, L. Xi, Q. Duan, J. Zhao, Rough set and ensemble learning based semi-supervised algorithm for text classification, Expert Systems with Applications. 38 (2011) 6300ï6306.
[32] M.A. Nielsen, Neural networks and deep learning, Determination press USA, 2015.
[33] A. Trask, A Neural Network in 11 lines of Python (Part 1) - i am trask, Https://Iamtrask.Github.Io. (2015). https://iamtrask.github.io//2015/07/12/basic-python-network/ (accessed May 20, 2018).
[34] A. Trask, A Neural Network in 13 lines of Python (Part 2 - Gradient Descent) - i am trask, Https://Iamtrask.Github.Io. (2015). https://iamtrask.github.io//2015/07/27/python-network-part2/ (accessed May 31, 2018).
[35] G. Kassabgi, Text Classification using Neural Networks, Machine Learnings. (2017). https://machinelearnings.co/text-classification-using-neural-networks-f5cd7b8765c6 (accessed March 21, 2018).
[36] B. Pang, L. Lee, A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts, in: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, 2004: p. 271. http://dl.acm.org/citation.cfm?id=1218990 (accessed November 28, 2014).
[37] J. Martineau, T. Finin, Delta TFIDF: An Improved Feature Space for Sentiment Analysis., Icwsm. 9 (2009) 106.

[38] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, C.-J. Lin, LIBLINEAR: A library for large linear classification, Journal of Machine Learning Research. 9 (2008) 1871ï1874.
[39] A.L. Maas, R.E. Daly, P.T. Pham, D. Huang, A.Y. Ng, C. Potts, Learning word vectors for sentiment analysis, in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1, Association for Computational Linguistics, 2011: pp. 142ï150.
[40] P.F. Brown, P.V. deSouza, R.L. Mercer, V.J.D. Pietra, J.C. Lai, Class-based N-gram Models of Natural Language, Comput. Linguist. 18 (1992) 467ï479.
[41] scikit-learn, scikit-learn: machine learning in Python, (2010). http://scikit-learn.org/stable/ (accessed April 4, 2018).
[42] S. Li, Scikit-Learn for Text Analysis of Amazon Fine Food Reviews, DataScience+. (2017). https://datascienceplus.com/scikit-learn-for-text-analysis-of-amazon-fine-food-reviews/ (accessed April 21, 2018).
[43] F. Pedregosa-Izquierdo, Feature extraction and supervised learning on fMRI: from practice to theory, PhD Thesis, Université Pierre et Marie Curie-Paris VI, 2015.
[44] J. McAuley, J. Leskovec, Hidden Factors and Hidden Topics: Understanding Rating Dimensions with Review Text, in: Proceedings of the 7th ACM Conference on Recommender Systems, ACM, New York, NY, USA, 2013: pp. 165ï172. doi:10.1145/2507157.2507163.
[45] M. Sokolova, G. Lapalme, A systematic analysis of performance measures for classification tasks, Information Processing & Management. 45 (2009) 427ï437.
[46] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Computation. 9 (1997) 1735ï1780.
[47] X. Zhang, J. Zhao, Y. LeCun, Character-level Convolutional Networks for Text Classification, in: Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1, MIT Press, Cambridge, MA, USA, 2015: pp. 649ï657. http://dl.acm.org/citation.cfm?id=2969239.2969312 (accessed July 10, 2018).
[48] S.J. Pan, Q. Yang, A Survey on Transfer Learning, IEEE Transactions on Knowledge and Data Engineering. 22 (2010) 1345ï1359. doi:10.1109/TKDE.2009.191.