Teleconsultations between Patients and Healthcare Professionals in Primary Care in Catalonia: the Evaluation of Text Classification Algorithms Using Machine Learning

Author	Affiliation	ORCID
Francesc López Seguí	TIC Salut Social - Ministry of Health, Barcelona, Spain CRES&CEXS –Pompeu Fabra University, Barcelona, Spain	0003-0977-0215
Ricardo Ander-Egg	Faculty of medicine - Barcelona University, Barcelona, Spain	0001-6971-7972
Gabriel de Maeztu	IOMED Medical Solutions, Barcelona, Spain	0003-0367-1822
Anna García-Altés	Agency for Healthcare Quality and Evaluation of Catalonia (AQuAS), Catalan Ministry of Health, Spain	0003-3889-5375
Francesc García Cuyàs	Sant Joan de Déu Hospital, Catalan Ministry of Health, Spain	0002-2448-5466
Sandra Walsh	Institut de Biologia Evolutiva (CSIC) – Pompeu Fabra University, Barcelona, Spain	0002-8761-4333
Marta Sagarra Castro	Centre d'Atenció Primària Capellades, Gerència Territorial de la Catalunya Central, Institut Català de la Salut, Sant Fruitós de Bages, Spain	0001-5219-6541
Josep Vidal-Alaball*	Health Promotion in Rural Areas Research Group, Gerència Territorial de la Catalunya Central, Institut Català de la Salut, Sant Fruitós de Bages, Spain Unitat de Suport a la Recerca de la Catalunya Central, Fundació Institut Universitari per a la recerca a l'Atenció Primària de Salut Jordi Gol i Gurina, Sant Fruitós de Bages, Spain Correspondence: jvidal.cc.ics@gencat.cat	0002-3527-4242

Abstract

Background: the primary care service in Catalonia has operated an asynchronous teleconsulting service between GPs and patients since 2015 (eConsulta), which has generated some 500,000 messages. New developments in big data analysis tools, particularly those involving natural language, can be used to accurately and systematically evaluate the impact of the service.

Objective: the study was intended to assess the predictive potential of eConsulta messages through different combinations of vector representation of text and machine learning algorithms and to evaluate their performance.

Methodology: 20 machine learning algorithms (based on 5 types of algorithms and 4 text representation techniques) were trained using a sample of 3,559 messages (169,102 words) corresponding to 2,268 teleconsultations (1.57 messages per teleconsultation) in order to predict the three variables of interest (avoiding the need for a face-to-face visit, increased demand and type of use of the teleconsultation). The performance of the various combinations was measured in terms of precision, sensitivity, F-value and the ROC curve.

Results: the best-trained algorithms are generally effective, proving themselves to be more robust when approximating the two binary variables "avoiding the need of a face-to-face visit" and "increased demand" (precision = 0.98 and 0.97, respectively) rather than the variable "type of query"(precision = 0.48).

Conclusion: to the best of our knowledge, this study is the first to investigate a machine learning strategy for text classification using primary care teleconsultation datasets. The study illustrates the possible capacities of text analysis using artificial intelligence. The development of a robust text classification tool could be feasible by validating it with more data, making it potentially more useful for decision support for health professionals.

Keywords: machine learning; teleconsultation; primary care; remote consultation; classification

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 December 2019 doi:10.20944/preprints201912.0220.v1

Peer-reviewed version available at *Int. J. Environ. Res. Public Health* **2020**, *17*, 1093; doi:10.3390/ijerph17031093

Introduction

eConsulta is an asynchronous teleconsultation service between patients and GPs as part of the electronic health records of the public primary healthcare system of Catalonia. In operation since the end of 2015, this secure messaging service was designed to complement face-to-face consultations with primary healthcare teams (PHT). It was gradually implemented up until 2017, when the service became available to every PHT: currently, all PHTs have used this tool at least once.

An earlier study analysed the reasons why patients sought a consultation which resulted in a patient-doctor interaction, as well as the subjective perception of the GP if they avoided a face-to-face visit or if it led to a consultation which otherwise would not have occurred, by means of a retrospective review of text messages relating to each case [Lopez 2019]. The results show there was a broad consensus among GPs that eConsulta has the potential to resolve patient queries (avoiding the need for a face-to-face visit in 88% of cases) for every type of consultation. In addition, GPs declared that ease of access led to an increase in demand (queries which otherwise would not have been made) in 28% of cases. Therefore, the possibility of eConsulta replacing a conventional appointment stands at between 88% and 63% [88% x (1-28%)]. The most common use of e-consultation was for the management of test results (35%), clinical enquiries (16%) and the management of repeat prescriptions (12%).

Technology offers new possibilities in conjunction with the aforementioned classical approaches. Artificial intelligence tools are already widely used in the field of healthcare in areas such as the prediction and management of depression, voice recognition for people with speech impediments, the detection of changes in the biopsychosocial status of patients with multiple morbidities, stress control, the treatment of phantom limb pain, smoking cessation, personalized nutrition by prediction of glycemic response, to try to detect signs of depression and in particular for reading medical images (Triantafyllidis 2019, Luo 2016, Li 2018, Gulshan 2016, Law 2019, Vidal-Alaball 2019).

The classification of texts in the medical field has also been used to conduct a review of influenza detection and prediction through social networking sites (Alessa, Xu, Doan, Heather) and in the analysis of texts from internet forums (McRoy, Bobicev). More specifically, in the framework of teleconsultations, a US-based study used machine learning to annotate 3,000 secure message threads involving patients with diabetes and clinical teams according to whether they contained patient-reported hypoglycaemia incidents (Chen 2019). As far as the authors are aware, no study has looked into the development of a text classification algorithm in the context of teleconsultations between patients and primary care physicians.

The present study aims to evaluate specific text classification algorithms for eConsulta messages and to validate their predictive potential. The algorithms have been trained using a vector representation of text from the body of the message and the three variable annotations that primary healthcare professionals in Central Catalonia used in a previous study: avoiding the need for a face-to-face visit, increased demand and type of use of the teleconsultation (López 2019). Our study represents an exhaustive exploratory analysis of text classification algorithms of teleconsultation messages between GPs and patients that can provide useful information for future research. and a potential use for decision support in healthcare.

Methodology

Data acquisition

The teleconsultations which had previously been classified that were used as the basis for training the algorithm are those which were acquired in the study by a previous study (López) (Table 1). They are part of the health records of the *Gerència Territorial de la Catalunya Central* of the *Institut Català de la Salut* covering the period from when the tool was first used until the date of its extraction for analysis purposes (8 April 2016 to 18 August 2018). Message deidentification was performed by substituting all possible names contained in the Statistical Institute of Catalonia database (IDESCAT 2019) with a common token and removing all other personal attributes. The classification method used for the conversations is described and justified by López et al. 2019: every healthcare professional who received an eConsulta labelled it according to whether, in their opinion, it avoided the need for a face-to-face consultation, led to an increased demand and by type of teleconsultation (Appendix 1). These results of this annotation, with the corresponding messages, were used to train the text classification model using the three variables previously mentioned (Table 2).

Table 1: Data recorded by the teleconsulting system.

Conversation title	Conversation ID	Message ID	From	To	Message	Files attached?
Travelling to Australia	C1	M1	Mr. John Patient	Ms. Jane Doctor	Dear doctor, I'm travelling to Australia on 15 December. Do I need to have any vaccinations? Many thanks	No
		M2	Ms. Jane Doctor	Mr. John Patient	Hi, Vaccinations are required for travel to Australia	No

Table 2: Annotation by the GP.

Conversation ID	Face-to-face visit avoided?	Increased demand?	Type of visit
C1	Yes	No	6 (Vaccinations)

Vector representation of text in eConsulta messages

The emails needed to be represented in some way in order to use them as input for the models. A common practice in machine learning is the vector representation of words. These vectors capture hidden information about the language, such as word analogies and semantics and improve the performance of text classifiers.

Four techniques have been used to generate the vector representation of texts. The Bag of Words (BoW) approach counts the number of times pairs of words appear in each document. The document is represented as a vector of a finite vocabulary. The Term-Frequency-Inverse Document Frequency (TF-IDF) method assigns paired words a weight depending on the number of times they appear in a particular document (the Term-Frequency), while discounting its frequency in other documents (Inverse Document Frequency): the more documents a word appears in, the less valuable that word is as a signal to differentiate any given document. Word2Vec is a two-layered neuronal network which trains and processes text. Its input is a corpus of text and its output is a set of vectors for the words in the corpus, with words represented by numbers. The initial vector assigned to a word cannot be used to accurately predict its context, meaning its components must be adjusted (trained) through the contexts in which they are found. In this way, repeating the process for each word, word vectors with similar contexts end up in nearby vector spaces. Fasttext [Bojanowski] is used to obtain word2vec vectors. Finally, the objective of Doc2vec is to create a numerical representation of a document, regardless of its length. This approach represents each document by a dense vector, which learns to predict the words in the document [Le]. In all cases, before carrying out the vectorization of the texts, these were first tokenized and any stop-words eliminated (those which are taken to have no meaning in their own right, such as articles, pronouns or prepositions).

In each instance, the vectors were enriched by supplementing them with similar texts in Catalan and Spanish [Ljubesic]. The external data used to enrich the corpus were models of interactions extracted from online databases with colloquial language similar to that used in eConsulta. Where augmented BOW, TF-IDF and Word2Vec were used, word and character length and word density were also used as predictor variables.

Training and testing AI algorithms

The task addressed in this study is a multiclass classification with respect to the type of visit and two binary classifications for the other two variables (avoiding visit and increased demand). For each text vector representation algorithm five different algorithms were implemented: Random Forest, Gradient Boosting (lightGBM), Fasttext, Multinomial Naive Bayes and Naive Bayes Complement [Rennie]. A convolutional neural network was also used using the augmented Word2vec vectors. We tested the performance of the algorithms through a stratified 10-fold cross-validation: during 10 iterations/trainings, 9 divisions served as learning and 1 as a test.

The coefficients of interest to evaluate the goodness of the algorithms were precision (the fraction of relevant instances between the retrieved instances/proportion of correct predictions of the total of all predicted cases) and sensitivity (the number of correct classifications for the positive class "true positive"). It was decided not to use the “accuracy” coefficient since it is a metric which, given an unbalanced dataset like the one under investigation, can result in a very high score in spite of the fact that the classifier works poorly, since it assesses the number of total hits without taking into account whether most of the data is of the same class. The F value is used to determine a weighted single value of accuracy and completeness. The diagnostic value is assessed by means of the ROC curve. The goodness of fit of all the coefficients is represented as a value between 0 and 1.

Python 3.7 and the following libraries were used for the algorithm training: numpy [Van der Walt], matplotlib [Hunter], seaborn [Waskom], altair [Altair], scikit-learn [Pedregosa], pandas [McKinney], gensim [Rehurek], nltk [Bird], fasttext [Bojanowski], pytorch [Paszke], lightGBM [Ke]. The majority of the code was carried out on Jupyter Notebooks [Kluyver].

Ethical considerations

The study was approved by the Ethical Committee for Clinical Research at the Foundation University Institute for Primary Health Care Research Jordi Gol and Gurina, registration number P19/096-P and carried out in accordance with the Declaration of Helsinki (WMA 2013).

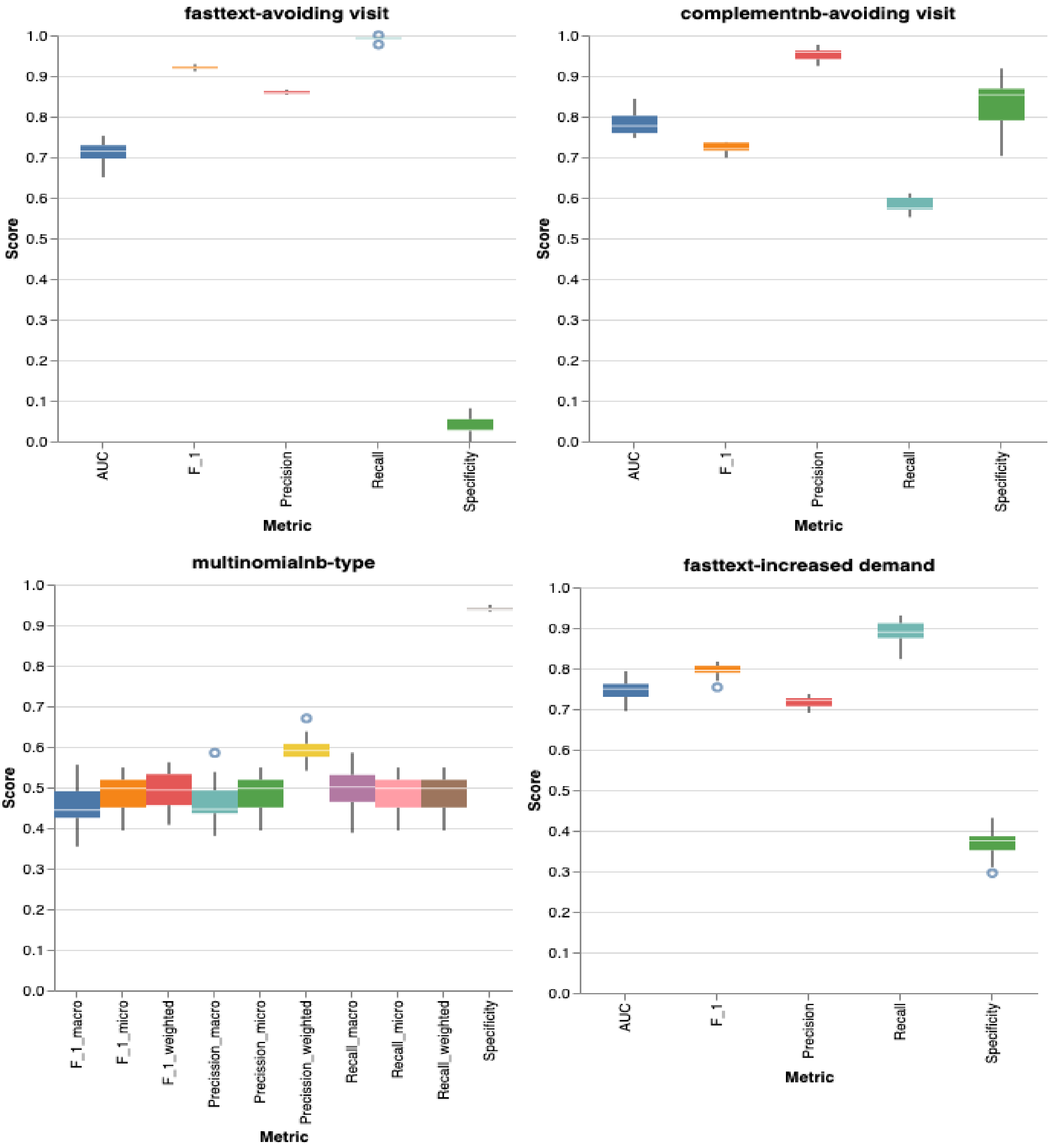
Results

In order to assess the predictive potential of eConsulta messages regarding the three variables of interest, we first aimed to identify the best combination of algorithms. A total of 3,559 messages (169,102 words) corresponding to 2,268 teleconsultations (1.57 messages per teleconsultation) were analysed in a framework of 20 different combinations of vector representation of text and machine learning algorithms (Table 3). We assessed the performance of the combinations of algorithms though a stratified 10-fold cross-validation analysis. Figure 1 shows the performance of the most stable algorithm (best metrics, in general) according to the predictor variable.

Table 3: Text representations and algorithms used

Text representations	Algorithms
BoW	Random Forest
TF-IDF	Gradient Boosting (lightGBM)
Word2Vec	Fasttext
Doc2Vec	Multinomial Naive Bayes
	Complement Naive Bayes

Figure 1: Performance metrics of algorithms.



Specific combinations of algorithms per variable generally perform very well. Table 4 shows the evaluation metrics (mean + standard deviation of the 10 iterations) of the combination of algorithm and numerical representation of the text which has a better performance for each target variable. For all of the cases, the vectors obtained directly from the original texts have been more useful than those enriched with external texts. Table 4 shows that algorithms are generally effective, showing they are better when approximating the two binary variables (avoiding the need for a face-to-face visit, increased demand) than the variable "type of query". Thus, eConsulta's classifiers have a promising and robust predictive value, especially for binary variables.

Table 4: Results of the best algorithm/text representation combination, according to the variable to be approximated. Average (SD) of the 10 iterations.

Variable	Precision	Recall	F1	Roc_AUC
Avoiding the need of a face-to-face visit	Random Forest TF-IDF 0.98 (0.026)	FastText Word2Vec 0.99 (0.005)	FastText Word2Vec 0.92 (0.004)	ComplementNB TF-IDF 0.79 (0.032)
Increased demand	Random Forest TF-IDF 0.97 (0.057)	FastText Word2Vec 0.89 (0.029)	FastText Word2Vec 0.79 (0.018)	FastText Word2Vec 0.75 (0.031)
Type of use of the teleconsultation (micro averaged score)	MultinomialNB BOW 0.48 (0.049)	MultinomialNB BOW 0.48 (0.049)	MultinomialNB BOW 0.48 (0.049)	

As a whole, the results illustrate eConsulta's algorithm classifiers potential predictive value and provide a valuable insight into the implementation of AI methodologies for healthcare teleconsultation.

Discussion

Although the study used all the available information, the major limitation of the analysis is the amount of data with which the algorithms were tested, meaning the conclusions must be understood in light of this shortcoming. This is especially relevant in the case of trying to calculate the variable "type of message", since the number of types which contain the classification (13) meaning the quantity of messages of each with which the classification algorithm has been trained is minimal, thus diminishing its predictive capacity. What is required is not only more messages, they must also contain as much information as possible. Validating the algorithm requires a replication of the proposed methodology with a larger data set, together with the analysis of subgroups. Likewise, the goodness of fit of the results may be caused by overfitting: the model explains this set of data well, but could show weaknesses when generalizing to others, limiting its potential for extrapolation. Because of that, this study includes exhaustive detail of the methodology used in order that it can be replicated.

Using complex mathematical models makes it difficult to explain why some work better than others. The vectors would need to be evaluated at a lower level in order to have a better idea as to which characteristics redirect the model towards one decision or another. This analysis is of interest for future applications of these techniques on a larger scale or for applications related to medical practice.

In Catalonia, the number of conversations and messages now stand at approximately 370,000 and 500,000 respectively. Applying a classification algorithm like the one proposed here would help us understand the nature of the conversations and their impact in real time. Future research should evaluate the use of automation (to send a diagnostic test, generate an alert or "thank you" and close the case) as a tool for decision support for healthcare professionals to improve the management of clinical cases and to save GPs time.

In summary, it has been established that the implementation of an algorithm for the prediction of factors such as a reduction in the number of face-to-face visits, induced demand or type of consultation is technically feasible and potentially useful in the context of service planning, management of the demand and evaluation. This study presents a combination of algorithms based on machine learning and more efficient representation vectors for this type of data. This study is an initial exploration into the potential and promising use of teleconsultation data.

References

1. López Seguí F, Vidal Alaball J, Sagarra Castro M, García Altés A, García Cuyàs F. Does teleconsultation reduce face to face visits? Evidence from the Catalan public primary care system. JMIR Preprints. 25/04/2019:14478
2. World Medical Association. WORLD MEDICAL ASSOCIATION DECLARATION OF HELSINKI. Ethical Principles for Medical Research Involving Human Subjects Helsinki 2013, https://www.wma.net/what-we-do/medical-ethics/declaration-of-helsinki/ (last accessed 6 September 2019).
3. Triantafyllidis AK, Tsanas A Applications of Machine Learning in Real-Life Digital Health Interventions: Review of the Literature J Med Internet Res 2019;21(4):e12286
4. Luo W, Phung D, Tran T, Gupta S, Rana S, Karmakar C, Shilton A, Yearwood J, Dimitrova N, Ho TB, Venkatesh S, Berk M Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research: A Multidisciplinary View J Med Internet Res 2016;18(12):e323
5. Li Z, Keel S, Liu C, He Y, Meng W, Scheetz J, et al. An Automated Grading System for Detection of Vision-Threatening Referable Diabetic Retinopathy on the Basis of Color Fundus Photographs. Diabetes Care. 2018;1–8.
6. Gulshan, Varun, et al. "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs." Jama 316.22 (2016): 2402-2410.
7. Vidal-Alaball, Josep, et al. "Artificial Intelligence for the Detection of Diabetic Retinopathy in Primary Care: Protocol for Algorithm Development." JMIR research protocols 8.2 (2019): e12539.
8. Alessa A, Faezipour M. Preliminary Flu Outbreak Prediction Using Twitter Posts Classification and Linear Regression With Historical Centers for Disease Control and Prevention Reports: Prediction Framework Study. JMIR Public Health Surveill 2019;5(2):e12383
9. Xu S, Markson C, Costello KL, Xing CY, Demissie K, Llanos AA. Leveraging Social Media to Promote Public Health Knowledge: Example of Cancer Awareness via Twitter. JMIR Public Health Surveill 2016;2(1):e17
10. Doan S, Ritchart A, Perry N, Chaparro JD, Conway M. How Do You #relax When You're #stressed? A Content Analysis and Infodemiology Study of Stress-Related Tweets. JMIR Public Health Surveill 2017;3(2):e35
11. McRoy S, Rastegar-Mojarad M, Wang Y, Ruddy KJ, Haddad TC, Liu H. Assessing Unmet Information Needs of Breast Cancer Survivors: Exploratory Study of Online Health Forums Using Text Classification and Retrieval. JMIR Cancer 2018;4(1):e10
12. Bobicev V, Sokolova M, El Emam K, Jafer Y, Dewar B, Jonker E, Matwin S. Can Anonymous Posters on Medical Forums be Reidentified? J Med Internet Res 2013;15(10):e215
13. Chen J, Lalor J, Liu W, Druhl E, Granillo E, Vimalananda VG, Yu H. Detecting Hypoglycemia Incidents Reported in Patients' Secure Messages: Using Cost-Sensitive Learning and Oversampling to Reduce Data Imbalance. J Med Internet Res 2019;21(3):e11990
14. IDESCAT. Noms de la població. Last accessed: 24 september 2019. http://www.idescat.cat/noms/
15. Bojanowski, Piotr, et al. "Enriching word vectors with subword information." *Transactions of the Association for Computational Linguistics* 5 (2017): 135-146.
16. Le, Quoc, and Tomas Mikolov. "Distributed representations of sentences and documents." *International conference on machine learning*. 2014.
17. Joulin, Armand, et al. "Bag of tricks for efficient text classification." *arXiv preprint arXiv:1607.01759* (2016).
18. Ljubeˇsic, Nikola, and Antonio Toral. "caWaC-A web corpus of Catalan and its application to language modeling and machine translation." *LREC*. 2014.
19. Rennie, Jason D., et al. "Tackling the poor assumptions of naive bayes text classifiers." *Proceedings of the 20th international conference on machine learning (ICML-03)*. 2003.
20. Batista, Gustavo EAPA, Ana LC Bazzan, and Maria Carolina Monard. "Balancing Training Data for Automated Annotation of Keywords: a Case Study." *WOB*. 2003.
21. Van Der Walt, Stefan, S. Chris Colbert, and Gael Varoquaux. "The NumPy array: a structure for efficient numerical computation." *Computing in Science & Engineering* 13.2 (2011): 22.
22. J. D. Hunter, "Matplotlib: A 2D Graphics Environment", Computing in Science & Engineering, vol. 9, no. 3, pp. 90-95, 2007.
23. Waskom, Michael, et al. "mwaskom/seaborn: v0. 9.0 (July 2018)." *DOI: https://doi. org/10.5281/zenodo* 1313201 (2018).
24. Pedregosa et al.,Scikit-learn: Machine Learning in Python, JMLR 12, pp. 2825-2830, 2011.
25. McKinney, Wes. "Data structures for statistical computing in python." *Proceedings of the 9th Python in Science Conference*. Vol. 445. 2010.
26. Rehurek, Radim, and Petr Sojka. "Software framework for topic modelling with large corpora." *In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. 2010.
27. Bird, Steven, Edward Loper and Ewan Klein (2009), Natural Language Processing with Python. O'Reilly Media Inc.
28. Paszke, Adam, et al. "Automatic differentiation in pytorch." (2017).
29. Ke, Guolin, et al. "Lightgbm: A highly efficient gradient boosting decision tree." *Advances in Neural Information Processing Systems*. 2017.
30. Kluyver, Thomas, et al. "Jupyter Notebooks-a publishing format for reproducible computational workflows." *ELPUB*. 2016.
31. Altair: https://altair-viz.github.io/index.html

Acknowledgements

This study was conducted with the support of the Secretary of Universities and Research of the Department of Business and Knowledge at the *Generalitat de Catalunya*.We would like to thank the staff at the Technical and Support Area of *Gerència Territorial de la Catalunya Central* for their assistance during the data collection phase.

Abbreviations

GP: General Practitioner
BOW: Bag of Words
TF-IDF: Term frequency – inverse document frequency
ROC: Receiver Operating Characteristics

Conflictsof interest

None declared.

Appendix 1: REASONS FOR USING eConsulta

ADMINISTRATIVE

1- Management of test results

o The patient provides the results of tests carried out in an external centre in order that they are recorded in their medical history
o The GP provides the results of tests with normal results
o The GP deals with questions related to tests requested by the patient
o The GP requests tests after conducting a follow-up teleconsultation

2- Temporary disability management

o The patient communicates changes to their health related to an upcoming temporary disability
o The GP tracks the progress of a temporary disability in conjunction with face-to-face visits

3- Management of visits/referrals
o The patient has an enquiry which the GP thinks ought to be dealt with by a specialist and refers them. They can also report incidents resulting from any referrals made
o The GP resolves incidents relating to the timing of visits
o The GP cancels visits from other clinicians in cases in which the problem has been resolved following completion of the e-consultation
o Validation of appointments with other specialists where the citizen needs more information about the motivation of the appointment

4- Request for a clinical report/sick-note

o The patient asks for a report/sick-note while consulting their medical history
o The GP asks the patient for more information in order to prepare the report

5- Repeat prescriptions

o The patient asks for their prescription to be updated if it has been modified by an external specialist, either because they do not use it or because it has expired
o The GP warns the patient that their prescription is about to expire and updates it
o The GP cancels an unnecessary prescription following an e-consultation

6- Vaccinations

o Updates of immunization schedules and general enquiries regarding vaccinations
o Questions concerning vaccinations for travel overseas

7- Other administrative issues: any administrative procedure which can be resolved without being physically present

MEDICAL

8- Medical enquiries: the patient has a question about their health that can be resolved without a physical examination. They can also attach photographs to accompany the description

9- Issues regarding medicines: the patient asks a question about a prescription

10- Questions regarding anticoagulants and dosage

OTHERS

11- Messages sent in error: the patient made a mistake

12- Other

13- Test messages