Intellectual Classifier Development of Citizens' messages on the “Our St. Petersburg” Portal: Experience in Using Machine Learning Methods

Petr Begen[0000-0002-0613-3133] and Andrei Chugunov[0000-0001-5911-529X]

[1] ITMO University, Kronverksky pr., 49, 197101, St. Petersburg, Russia
peetabegen@yandex.ru, chugunov@itmo.ru

Abstract. Functional features are investigated and shortcomings in the existing process of sending messages about city problems on the “Our St. Petersburg” portal are revealed. The approach to automatic classification development of citizens ' messages by existing on portal categories is described. Based on reports submitted by citizens in the amount of 1.5 million, training and test samples were formed in the ratio of 80% and 20% of texts main volume, respectively. Based on training data sample and 194 categories, the algorithm of automatic classification was trained using such classical methods of machine learning as naive Bayes classifier, decision trees and artificial neural networks. Using the method of determining effectiveness of the classification and test sample, trained algorithm was tested and checked. The analysis revealed that algorithm based on the use of artificial neural networks shows the best result among the other methods used. The average classification accuracy of the algorithm was approximately 82%. The trained algorithm was used in the development of an intellectual classifier, which is a web application and implements API mechanisms for interaction with main modules of the portal information system.

Keywords: Artificial Intelligence, Machine Learning, Artificial Neural Networks, Classifier, e-participation.

1 Introduction

Nowadays information technology usage for improving public administration has ceased to be perceived as a kind of innovation of technologies and systems of e-government have already entered everyday life of citizens and have become an integral part of state machine. Research and development are carried out not in the field of translation traditional organizational processes into electronic form, but in the field of improving information systems’ efficiency. The research presented in this article refers to this type of development.

Recently, more and more attempts are being made to formulate criteria for e-governance and e-participation effectiveness as a mechanism for feedback from governments to citizens. Improved information systems according to researchers is an important factor in the growth of institutional citizens’ trust to actions of the authorities and opportunities to influence these actions [1], and government responsiveness for e-citizens – the basic criterion of e-participation efficiency [2].

In 2018 Russian Federation was developing approaches to the restructuring policy state in the field of innovative development and informatization. This was due to the designation of a new priority and the adoption of Russian Federation national program "Digital economy". These processes stimulated a new round of interest in problems of optimization and improvement of state information systems' work, including problems of increasing their functioning efficiency. Currently, quite a lot of processes in e-governance systems require participation of government officials or subordinate organizations, and tasks' implementation involving the automation of individual operations can significantly improve efficiency of state information systems.

One of the approaches that began to be used in the development of state information systems is to use "Artificial intelligence" (AI), which is considered as one of the main trends in the modern information technologies (IT) development and is included in all lists of so-called "breakthrough technologies". There are many publications of an analytical and prognostic plan in which AI technologies play a key role at present stage of digital transformations [3], which is often referred to the "Industry 4.0" development.

It should be noted that interest on the part of developed countries in forming focused approach to AI development and ensuring introduction of these technologies and methods began in 2017–2018. At this time, countries such as Canada, China, Denmark, Finland, France, India, Italy, Japan, Singapore, South Korea, Sweden, Taiwan, the UAE and the UK adopted strategic documents to promote the development and use of AI [4]. In these documents, such areas as research, development of education system, AI usage promotion in the public and private sectors, ethics and legal aspects of application, standards and data infrastructure protection of digital information are identified in different degrees of elaboration.

Present work is a local research project under the direction oriented to the study of information systems' functioning specifics supporting electronic interaction of citizens with authorities in a variety of contexts: from applied research to create models of e-governance institutional environment functioning.

Within the framework of this research direction a series of projects is being implemented devoted to the empirical analysis of e-Participation practices, which is defined as "a set of methods and tools that ensure electronic interaction between citizens and authorities in order to take into account the views of citizens in state and municipal administration when making political and managerial decisions" [5, p. 60]. The pilot project, results of which are presented in this paper, is aimed at solving the problem of automating messages classification posted by citizens on the "Our St. Petersburg" portal. Using "Our St. Petersburg" portal residents of the city can send messages related to housing and communal services and city improvement, the state of sidewalks and roads, get background information on the object of interest of the city economy, etc. An important component of the portal is a system of organizational measures and rules of processing messages which involves many services and authorities of St. Petersburg [6].

2 Functional Features of the “Our St. Petersburg” Portal

2.1 About the “Our St. Petersburg” portal

The “Our St. Petersburg” portal was created on the initiative of the St. Petersburg Governor Poltavchenko G. S. in 2014 for the operational interaction between residents with representatives of St. Petersburg. Using the portal user has following opportunities [7]:

- to send messages on problems connected with housing and communal services and improvement of the city, a condition of roads and sidewalks, illegal objects of construction and trade, violation of the land or migratory legislations;
- to inform the city about the lack of background information on the Bulletin boards, also about the poor sanitary condition of the premises in the budgetary institutions of education, health, culture, social protection, employment;
- to receive additional information concerning the address city programs and managing organizations, also reference information on the object of municipal economy interest;
- get acquainted with technical and economic passports of apartment buildings in St. Petersburg and get information about the organizations that provide their service.

Messages sent through the “Our St. Petersburg” portal without fail are considered by city services in strictly established terms depending on chosen category according to the messages’ classifier. The portal user has the opportunity to receive information about the progress of consideration and processing of sent messages as well as to evaluate the response received.

The first version of the “Our St. Petersburg” portal was opened in January 2014 and during the year was carried out a gradual modernization and development of regulations for processing citizens ' appeals. From the very beginning it was decided to develop this information resource based on the City monitoring center, which processes telephone calls of citizens (operational services of the city on various problems).

Up to date there is a rapid development of the portal: as of December 2019, citizens of St. Petersburg filed more than 2 million messages about urban problems and the same number has been resolved (more than 96% of the messages submitted), and the number of registered users is about 157 thousand people and still the current figure continues to grow.

As of December 2019, the scheme of functional relations of the “Our St. Petersburg” portal, presented in [6], is as follows (Fig. 1).

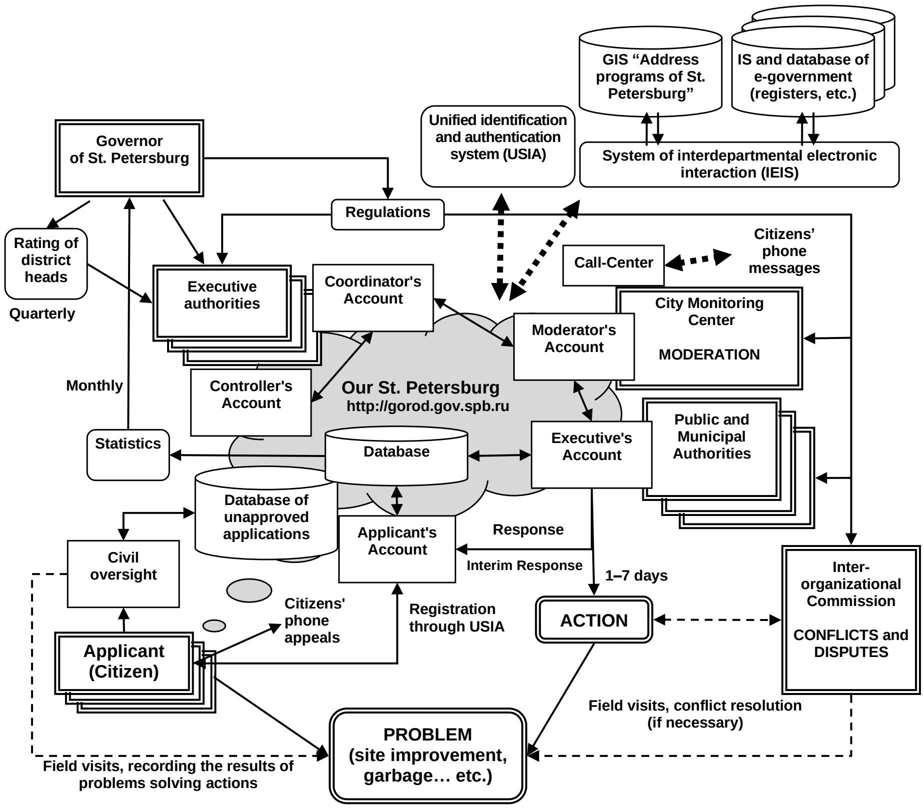
GIS "Address programs of St. Petersburg"
IS and database of e-government (registers, etc.)
Unified identification and authentication system (USIA)
System of interdepartmental electronic interaction (IEIS)
Governor of St. Petersburg
Regulations
Rating of district heads
Quarterly
Citizens' phone messages
Call-Center
Executive authorities
Coordinator's Account
Moderator's Account
City Monitoring Center
MODERATION
Monthly
Controller's Account
Our St. Petersburg
http://gorod.gov.spb.ru
Public and Municipal Authorities
Statistics
Database
Executive's Account
Database of unapproved applications
Applicant's Account
Response
Interim Response
Civil oversight
1–7 days
Inter-organizational Commission
CONFLICTS and DISPUTES
Citizens' phone appeals
Registration through USIA
ACTION
Applicant (Citizen)
Field visits, conflict resolution (if necessary)
PROBLEM (site improvement, garbage… etc.)
Field visits, recording the results of problems solving actions

Fig. 1. Information and institutional architecture of the "Our St. Petersburg" portal (2019)

For more than 5 years of portal existence (2014–2019) the following indicators were achieved:

- the number of problem categories increased 3.5 times: from 56 to 194;
- the number of authorities involved in the work increased 2.4 times: from 23 to 56;
- the number of organizations/performers increased 10 times: from 54 to 540;
- the number of the applicants' personal accounts have increased 4.8 times: from 6,344 to 30,640;
- the number of registered users on the portal has increased more than 9 times: from 17 thousand to 157 thousand;
- the number of initial problems reports has increased by an average of more than 140 times.

According to these statistics we can conclude that a sufficiently large and growing popularity of the portal among the city residents. Thus, the "Our St. Petersburg" portal becomes one of the most effective tools of e-participation implemented in St. Petersburg, thanks to publicity and openness, many problems were quickly brought to the city administration and successfully solved in the shortest possible time. In this regard, the load on the portal increases significantly: according to statistics [8], presented by the Administration of St. Petersburg, the “Our St. Petersburg” portal receives up to 2.5 thousand messages from citizens every day and about the same number of responses from Executives, and such a high load on moderating services (which now has 22 moderators in team) leads to problems of effective activities of various service types and raises the question of optimizing the existing functionality of the portal for faster processing of incoming citizens’ messages and their further transfer to Executive authorities.

In this paper the process of submitting a message to the portal and its classification in accordance with requirements will be considered in more detail and a solution for optimizing this process will be presented further.

2.2 The Current Process of Submitting Message to the Portal

Existing process of messages’ submission by the user to the portal is arranged as follows:

1. Firstly, you must log in and register on the portal. This is possible with the help of an account in social network “Vkontakte”, through the Unified identification and authentication system (USIA) or through the Unified single sign-in system.
2. To report a problem, you must select one of the 194 available problem categories using standard keyword search form, which will prompt several options for the desired category based on the query.
3. Specify the location of problem on the map by adding a point on the map or by entering street name and house number in the search bar.
4. Add photo of any supported format, confirming presence of the problem.
5. Briefly describe your problem in a special window (up to 1000 printed characters). When describing problem avoid abbreviations, obscene language, messages, requests, petitions of a personal nature.
6. If necessary, specify the name of the organization to which the user has already applied previously (* if necessary).
7. Confirm sending the generated message about the existing problem by pressing the “Send message” button.

In the considered process of sending a message in the 2nd item of list found a significant drawback, which is quite difficult to determine the correct category of messages for user from a large amount of proposed number. According to statistics moderators reject 20–25% of incoming citizens’ messages due to the discrepancy of the message about problem of the one of available categories proposed in the classifier. Thus, the proposed solution to optimize this activity should simplify the process of submitting a message to the user, speed up the process of checking the message for compliance with the requirements for moderating services and reduce the percentage of messages rejection due to an incorrectly selected category of problem.

2.3 An Approach to Process Optimization of Classifying Messages when they are Submitted to the “Our St. Petersburg” Portal

As been already mentioned earlier the moderation service, which now has 22 moderators, within one working day must work out each received message in accordance with the Order of work with messages. In days of peak loads the number of messages grows in one and a half or two times.

Based on these statistics, we can calculate the approximate time to work off each incoming message to the portal. Under condition officially adopted 8-hour working days everyone the moderator needs fulfill almost 114 new messages from users, thus on effective practicing 1 messages the moderator can spend no more than 4 minutes working time. According to our calculations, it takes up to 1–1.5 minutes for the moderator to determine whether the category corresponds to the declared problem described in the text of the message. In day’s peak leverage, when number of incoming messages expands in 1.5–2 times, number of messages, which every moderator should fulfill in for working days grows to 170–228, and time on practicing every message is shrinking until 2–3 minutes. Therefore, the terms of moderation in such emergency situations have to be increased, which is promptly reported in the “news” section on the portal. As a result of implementation of the decision on process optimization of message classification it is planned that the time allocated for check of text conformity of the message to the set category and requirements will be considerably reduced and will make no more than 30 sec further.

Also, in statistics it is specified that 20–25% of all arriving messages from citizens are rejected by moderators because of incorrectly chosen category. Thus, in order to optimize this process, it is necessary to form certain criteria under which it will be possible to increase the efficiency of this algorithm and reduce the percentage of messages rejection up to 15%.

As one of the solutions, it is proposed to develop and implement automatic classification of citizens ' messages. In order to minimize the risk of erroneous definition of the category by the user and to increase the efficiency of moderating service for processing incoming messages, the following approach to solving this problem is proposed:

- to submit a problem report it is necessary to exclude the obligation for the user to choose a problem category from the Classifier on his own or enter keywords in the search form: for this purpose, the user needs just to describe the problem in the form of a message text. Follow the procedure when submitting, such as specifying the location of an existing problem on a map and uploading supporting photos, save.
- for moderating service to develop the module of automatic text message classification which will present result of work in the form of the ranked list from three certain categories with the corresponding percent of classification accuracy for the subsequent choice by the moderator.

Methods for implementing this approach are described further.

3 Intellectual Message Classifier

3.1 Algorithm of messages' automatic classification

AI technologies such as machine learning and Natural Language Processing techniques have been proposed to implement automatic message text classification.

To achieve stated goal following tasks were set:

- prepare data for training and testing classification algorithm;
- apply to obtained data basic methods of Natural Language Processing;
- build and train a classification model based on machine learning techniques;
- test trained model on the basis of a test sample and get accuracy score for further analysis of result.

Data of citizens' messages were obtained from portal database in amount of 1.5 million. When sending a message to the portal it already has a category that is defined by user himself, so messages checked and accepted by the moderating service were used as data. In accordance with common practice data were divided into training and test samples in a ratio of 80/20. Note that test sample does not participate in training of the model, which means that the model will “see” this data for the first-time during testing. This approach allows us to obtain objective estimates of trained model classification accuracy.

For the model to be able to work with incoming data stream, it is necessary to preprocess and represent it in numerical form. At preparatory stage all obtained data is processed: punctuation marks, invisible symbols and numbers are removed, words are converted to lower case and initial form (for words with different prefixes, suffixes and endings) [9].

TF-IDF measure [10] is used to represent an array of data in the form of numeric vectors, which reflects importance of using each word from a certain set of words (number of words in the set determines the dimension of vector) in each body of text. Also, technique helps to exclude the most frequently encountered words (for example, prepositions and conjunctions) or Vice versa rarely encountered, because such words carry little useful information and only add information noise to unstructured text bodies.

Another point of improving search for significant features in the text was formation of a stop-words list, which mainly includes names of streets or urban facilities, also do not have a significant impact on the definition of problem category.

As another Natural Language Processing method, Word2Vec algorithms [11] were used to represent words in vector space. Algorithms use text context to form numerical representations of words, so words used in the same context have similar vectors. This approach also provides an effective way to identify significant features in text to improve the final result of classification.

To build a classification model based on analysis of works the following machine learning methods were chosen, showing good results when working with text information: naive Bayesian classifier [12], decision tree [13] and artificial neural networks [14]. Three networks with different architectures have been proposed as neural networks: feed-forward network (FFN), convolutional network (CNN) and recurrent network (RNN) with LSTM block. Each of methods organizing neural networks architecture has its advantages and disadvantages, but each has good results in classification problems, so it was decided to apply different methods and architectures and analyze the result within conditions of our problem.

The model was developed with Python programming language. Keras framework (with an add-on over TensorFlow mechanisms) and scikit-learn library were used to implement machine learning methods and configure neural network architectures.

After training model with different methods tests were conducted since a test sample. To assess quality of trained model we used metric F-measure, which is the harmonic average between Precision and Recall of classification. The common formula of metric F-measure has the following form [15]:

$$F\text{-}measure = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{1}$$

Precision is a proportion of true texts belonging to a given category relative to all texts that the model has assigned to that category, and Recall is a proportion of texts found by the model belonging to category relative to all texts of that category in the test sample.

The results of trained models with different machine learning methods are presented in the Table 1 below:

Table 1. F-measure and training time for different machine learning methods

	Machine learning method's name				
	Naive Bayesian classifier	Decision tree	FFN	CNN	RNN with LSTM
F-measure	0.659	0.703	0.7836	0.8199	0.8095
Model’s training time	20 min	5 h 14 min	1 h 45 min	2 h 5 min	2 h 35 min

According to analysis of presented F-measure accuracies indicators the best and quite fast learning method machine learning, which was applicable in our classification tasks, was convolutional neural network (CNN), which showed almost 82% of accuracy in identifying category of problem based on the body of text message. The model with a recurrent neural network (RNN) with an LSTM block, which is traditionally one of the best in text classification problems nowadays, performed slightly worse (i.e. a difference of 1%). Thus, an algorithm using a convolutional neural network as one of the best performed was proposed in further development of intellectual message classifier.

3.2 Criteria for the Success of Intellectual Message Classifier Operation

In order to implement this business function in the existing functionality of the system and ensure the correctness of the automatic classification it is necessary to create a list of criteria.

As a result of the algorithm analysis for working with messages and business functions a final list of criteria for the success of the tools for selecting the subject of citizens' messages and optimizing the process of submitting a message to the portal was compiled which looks as follows:

- the subject of the message must correspond to the categories and time period for the occurrence or elimination of problems specified in the Classifier;
- the message should contain a description of the problem in only one of the categories;
- the text of the message (if necessary) should contain the same coordinates of the problem location as the coordinates corresponding to the selected location on the map;
- the message should not coincide with other message (on set of parameters: object, category, reason, address / coordinates of a problem) which is placed on the portal and is under consideration;
- the message should not contain groundless, unproven charges against Executive bodies of the St. Petersburg state power and the state institutions (enterprises) subordinated to them, Federal bodies of the state power, physical persons or legal entities;
- the message should not contain personal data of third parties distributed without their consent;
- the message should not contain messages, requests, petitions of a personal nature related to the work of the portal;
- the message should not contain information distributed for commercial purposes or for any other purposes other than the purposes of the Order (including spam, advertising in the message text, images, video files, links to third-party resources of the information and telecommunication network “Internet”);
- the message must be a logically complete statement, not contain typos and (or) errors that prevent the understanding of the meaning of the appeal or allow for its ambiguous interpretation;
- the message must contain a stylistically correct request, corresponding to the norms of business communication;
- the message should be written in Cyrillic preferably in lowercase letters, not contain inappropriate abbreviations and obscene language;
- the text of the message should not exceed the limit of 1000 characters;
- in the Classifier it is necessary to exclude possibility of duplication of categories, texts of messages;
- for each category there must be at least 30 examples of relevant message text to successfully train the classification model;

- the percentage of accuracy in determining each category using machine learning methods should not be less than 80% and constantly improve.

In compliance with the formed criteria for the success of the automatic classification, it is planned to significantly reduce the average time for working off 1 message from the moderator by at least 25% (from 4 minutes to 3 under normal load on the portal), reduce the percentage of messages rejection due to an incorrectly selected category (from 20–25% to 15–20%, i.e. by at least 5%), improve the usability of the portal and facilitate the process of submitting a message to the user.

The intellectual message classifier is designed for moderating services in order to improve efficiency and convenience of working with citizens' messages and is going to be a web application that implements API mechanisms for interaction with existing modules of information system of the “Our St. Petersburg” portal.

The developing classifier will allow to automatically determine category of the user's message in asynchronous mode and present the result for moderating services in the form of a ranked list of three most possible categories with an indication of definition accuracy percentage. If the definition percentage of any category is below 5%, then submitted message does not match any of the available categories, which so will also prompt the services to make a further decision. This approach will allow services to accurately verify correctness of problem category choice proposed by classifier as well as faster to consider text message at the time of detecting possible errors and passing it on to Executive authorities.

4 Conclusions

As a result of this work the functional interaction of the “Our St. Petersburg” portal's components was considered, the processes of submitting a message to the portal and its further development by various services were described.

Based on the identified problems that negatively affect the effective operation of moderating and other services of the portal when working with citizens' messages an approach to the solution of the optimization process related to the messages' classification was proposed.

At this stage an algorithm was developed for automatic classification of citizens' messages into categories on the “Our St. Petersburg” portal based on machine learning methods. The algorithm was trained on data previously divided into training and test samples in a ratio of 80/20, respectively, as well as analyzed and presented in vector space using Natural Language Processing methods.

The best machine learning method used in automatic classification algorithm was the convolutional neural network (CNN) which showed an average category determination accuracy (i.e. F-measure) of about 82%. The developed algorithm with this method was used in further development of an intellectual classifier for moderating services.

As a further stages it is planned to explore the use of intellectual classifier in the framework of the tasks for the compliance of communications approved the rules according to the Order of messages in automatic mode and analysis to identify the increase of services activity efficiency of the portal.

This work was supported by the Russian Science Foundation, project No. 18-18-00360 “E-participation as Politics and Public Policy Dynamic Factor”.

References

1. Jansen, A.: The understanding of ICTs in public sector and its impact on governance. In: Electronic government: Proceedings of the 11th IFIP WG 8.5 international conference EGOV-2012, LNCS book series, vol. 7443, pp. 174–186 (2012).
2. Vidyasova, L.A., Misnikov, Y.G.: Kriterii ocenki social'noj effektivnosti portalov elektronnogo uchastiya v Rossii. Informacionnye resursy Rossii 5(159), 16–19 (2017).
3. Pandya, J.: The Geopolitics of Artificial Intelligence, https://www.forbes.com/sites/ cognitiveworld/2019/01/28/the-geopolitics-of-artificial-intelligence/#5a4b420979e1, last accessed 2019/12/07.
4. Dutton, T.: An Overview of National AI Strategies, https://medium.com/politics-ai/an-overview-of-national-ai-strategies-2a70ec6edfd, last accessed 2019/12/07.
5. Chugunov, A.V.: Vzaimodejstvie grazhdan s vlast'yu kak kanal obratnoj svyazi v institucional'noj srede elektronnogo uchastiya. Vlast' (10), 59–66 (2017).
6. Chugunov, A.V., Rybalchenko, P.A.: Razvitie sistemy elektronnogo vzaimodejstviya grazhdan s vlastyami v Sankt-Peterburge: opyt portala “Nash Peterburg”: 2014–2018 gg. Informacionnye resursy Rossii (6), 27–34 (2018).
7. O portale, https://gorod.gov.spb.ru/about/, last accessed 2019/12/09.
8. Portalu “Nash Sankt-Peterburg” – pyat’!, https://www.gov.spb.ru/gov/otrasl/ c_information/news/159410/, last accessed 2019/12/08.
9. Zibert, A.O., Hrustalev, V.I.: Razrabotka sistemy opredeleniya nalichiya zaimstvovanij v rabotah studentov vysshih uchebnyh zavedenij. Metody predvaritel'noj obrabotki teksta. Universum: Tekhnicheskie nauki: elektron. nauchn. zhurn 4(5), (2014), http://7universum.com/ru/tech/archive/item/1258, last accessed 2019/12/07.
10. Ingersoll, G.S., Morton, T.S., Ferris, E.L.: Obrabotka nestrukturirovannyh tekstov. Poisk, organizaciya i manipulirovanie. Per. s angl. Slinkin, A.A. DMK Press, Moscow (2015).
11. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient Estimation of Word Representations in Vector Space. (2013), https://arxiv.org/abs/1301.3781, last accessed 2019/12/06.
12. Barsegyan, A.A., Kupriyanov, M.S., Holod I.I., Tess M.D., Elizarov S.I.: Analiz dannyh i processov: ucheb. Posobie. 3d izd., pererab. i dop. BHV-Peterburg, St. Petersburg (2009).
13. Aggarwal, C.C.: Data Classification: Algorithms and Applications. 1st edn. Chapman & Hall/CRC (2014).
14. Prasanna P.L., Rao, D.R.: Text classification using artificial neural networks. International Journal of Engineering & Technology 7 (1.1), 603–606 (2018).
15. Sasaki, Y.: The truth of the F-measure. Teach Tutor Mater 1(5), 1–5 (2007).