

TecNM at MEX-A3T 2020: Fake News and Aggressiveness Analysis in Mexican Spanish

Samuel Arce-Cardenas^a, Daniel Fajardo-Delgado^a and Miguel Á. Álvarez-Carmona^{b,c}

^a*Tecnológico Nacional de México / Campus Ciudad Guzmán, Mexico.*

^b*Centro de Investigación Científica y de Educación Superior de Ensenada, Mexico*

^c*Consejo Nacional de Ciencia y Tecnología (CONACYT), Mexico*

Abstract

This paper describes our participation in the MEX-A3T 2020 for the tasks of identification of aggressiveness and fake news in Mexican Spanish tweets. We evaluate the combination of basic text classification techniques, including six machine learning algorithms, two methods for keyword extractions, and two preprocessing techniques. Our best run showed an F1-macro score of 0.754 for aggressiveness and 0.815 for fake news. Our preliminary results are satisfactory and competitive with other participating teams.

Keywords

Aggressiveness Identification, Fake News Classification, Natural Language Processing

1. Introduction

In today's digital culture, people spend more time on online social networks as a medium to interact, share, and collaborate with others using a style of informal communication [1]. However, these social networks are not exempt from unappropriated conducts and misbehaviors intended to cause emotional pain or to harm society through the communication process [2]. One of these destructive features of communications is the aggressiveness, a trait that involves attacking the self-concept of others [3]. The other one lies in the threat of disinformation, designed to negatively influence people and provide them an incorrect insight into different situations [4]. Both of these problems are tasks covered on the MEX-A3T 2020, a forum designed to encourage research on the analysis of social media content in Mexican Spanish [5][6].

In this work, we approach the tasks of aggressiveness and fake news posed by the MEX-A3T 2020 from a machine-learning perspective. Each of the tasks represents a binary classification problem for text content written in Mexican Spanish. The corpus for the aggressiveness task consists of 6593 tweet feeds geolocated in Mexico City. On the other hand, the corpus for the fake news task consists of 637 texts collected from January to July of 2018 from newspaper websites, media companies, and other particular websites. This work is motivated to evaluate when using basic text classification techniques is enough to provide competitive results.

Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020)

EMAIL: samuel11290806@itcg.edu.mx (S. Arce-Cardenas); dfajardo@itcg.edu.mx (D. Fajardo-Delgado);


malvarezc@cicese.mx (M.Á. Álvarez-Carmona)

ORCID: 0000-0002-2547-0047 (S. Arce-Cardenas); 0000-0001-8215-5927 (D. Fajardo-Delgado); 0000-0003-4421-5575

(M.Á. Álvarez-Carmona)



© 2020 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

2. State of the art

The MEX-A3T is an evaluation forum for IberLEF intended for the research in natural language processing (NLP) and considering a variety of Mexican Spanish cultural traits. In this vein, the 2018 edition was the first to consider the aggressiveness identification for Mexican Spanish tweets [7]. The winning team for the aggressiveness task for that edition was INGEOTEC [8], obtaining an F1-macro score of 0.620. Another interesting result was the development of linguistic generalization of the typical Mexican slang used in tweets to reduce the impact of size on the word bag [9]. For the 2019 edition of the MEX-A3T track [10], the approach of the University of Chihuahua (UACH) [11] obtained the best performance, outperforming all proposed baselines, except the results from the winner team of the 2018 edition. Nevertheless, the UACH approach is considerably much simpler than the one from INGEOTEC.

On the other hand, there are few studies on the detection of fakenews in Spanish [12] [6], one of these studies evaluates the complexity, the stylometric and psychological characteristics of the text in a multilingual setting [12], they used corpus of news written in American English, Brazilian Portuguese and Spanish, they used four classifiers, k-Nearest Neighbors, Support Vector Machine, Random Forest, and Extreme Gradient Boosting, and obtained an average detection accuracy of 85.3% with Random Forest. Another interesting investigation in which they created a new corpus of news in Spanish [6], with the true and fake tags used for automatic detection of fakenews, and presenting a fakenews detection method based on algorithms of classification of lexical characteristics such as Bag of Words, part of speech tag, n-grams (with n ranging from 3 to 5) and the combination of n-grams, the best result they obtained with an accuracy of 76.94%.

3. Methodology

The methodology of this work consists of the following steps: text preprocessing, text representation, and the building of the classification models.

Text preprocessing is commonly the first step in the pipeline of an NLP system, and it includes a set of techniques designed to transform text documents into a suitable representation form for automatic processing. The preprocessing techniques we employed in this work included the use of regular expressions, the tokenization, the deletion of punctuation, symbols, stop words, and the stemming. The regular expressions allowed us to identify some incorrect words for the Mexican Spanish, mainly those in which the same vowel appears subsequently three times or more. The best way to do this was by employing the "re" library in Python.

We also used the natural language toolkit (NLTK) to perform the tokenization, breaking the texts into words as essential elements. During this process, we also removed the punctuation marks, the special characters or symbols, as well as unnecessary stop words such as "el", "la", "los". Afterward, we used the Snowball stem library to reduce derived words into their original form or stem by performing the truncation of suffixes. Finally, to reduce even more the number of unmeaningful words, we ignored those that appear less than 20 or 40 times.

After the text preprocessing, we intended to identify the set of words that best describe the textual context. Extracting these words, also called terms or keywords, is the process to assign

a numerical value that represents the relevance of each word concerning the others within the corpus. In particular, we used two methods based on a simple statistic approach, the term frequency (TF), and the term frequency-inverse document frequency (TF-IDF). TF defines the local importance that each term has in a document based on its frequency; i.e., if a word w frequently appears in a document, then more important is w . IDF captures how many documents a word appears concerning the total number of words in the corpus, i.e., it highlights the rarity of the word. We used the implementations of TF and TF-IDF included in the scikit-learn library.

Finally, in order to build the classification models, we used the following machine learning algorithms implemented in scikit-learn: the k -nearest neighbors (KNN) for $k = 3, 7, 11$, the support vector machine (SVM) with a linear and a radial basis function (RBF) kernels, Decision trees (DT), Neural net (NN), and Naive Bayes (NB). We generated these models using the training set by using 10-fold cross-validation.

4. Experimental results

We divided the data set into 10 taking the first subset as validation and the other subsets as training, and we obtained the confusion matrix, then we take the second subset as validation and the rest as training we repeat this process until each subset has been into the validation set. Finally, we added the confusion matrices, and from this, we get the presented results.

Tables 1 and 2 show the performance of the proposed classification models applied to the fake news data set by using the TF and TF-IDF methods, respectively. The best result for this data set is by the combination of NN without using the techniques of stop words and stemming, and regardless of the use of TF and TF-IDF. Note that, except for the SVM with RBF, there is a notable difference between the results of NN concerning the rest. Also note that, in general, the results are slightly better when using TF-IDF than TF.

On the other hand, Tables 3 and 4 show the performance of the proposed classification models applied to the aggressiveness data set by using the TF and TF-IDF methods, respectively. The best result for this data set is by the combination of NN with the TF-IDF method and without using the techniques of stop words and stemming. Like the fake news data set, the results for the aggressiveness data set are slightly better when using TF-IDF than TF. On the other hand, and unlike the fake news classification results, the best model by using the TF method is the SVM with RBF. All of these results were obtained by ignoring the words that are repeated less than 20 times for both of the data sets (Tables 1-4). We omitted to report the results for the case when we ignored the words repeated less than 40 times. This because of the poor results and space limitations in the paper. On the other hand, the fake new data set includes, in addition to the complete text of the news, a header that describes the title of the news. We performed experiments either considering the header and not considering it. Tables 1 and 2 show only the results when the header is not considered, since these present better results.

Finally, for both of the data sets, the best results were obtained by preserving the stop words and omitting the steaming process. We conjecture that considering such words for these particular cases may distinguish the classes (aggressiveness/fake news) in the texts.

Table 1

Performance results of the proposed models for the FakeNews dataset by using TF in the validation stage.

| Classifier | Accuracy | Precision | Recall | F-measure | Stopwords | Stemming |
|------------|---------------|--------------------|--------------------|--------------------|-----------|----------|
| KNN_3 | 63.265 | 0.664±0.069 | 0.631±0.223 | 0.613±0.088 | Yes | Yes |
| KNN_7 | 62.48 | 0.668±0.083 | 0.623±0.259 | 0.597±0.106 | Yes | Yes |
| KNN_11 | 61.538 | 0.660±0.082 | 0.613±0.268 | 0.584±0.114 | Yes | Yes |
| L_SVM | 50.392 | 0.252±0.252 | 0.500±0.500 | 0.335±0.335 | Yes | Yes |
| RBF SVM | 74.411 | 0.744±0.004 | 0.744±0.003 | 0.744±0.001 | Yes | Yes |
| DT | 65.62 | 0.657±0.012 | 0.656±0.027 | 0.656±0.008 | Yes | Yes |
| NN | 76.766 | 0.768±0.000 | 0.768±0.005 | 0.768±0.003 | Yes | Yes |
| NB | 65.777 | 0.658±0.005 | 0.658±0.004 | 0.658±0.001 | Yes | Yes |
| KNN_3 | 62.48 | 0.670±0.085 | 0.623±0.262 | 0.596±0.108 | No | Yes |
| KNN_7 | 63.108 | 0.699±0.115 | 0.629±0.296 | 0.594±0.122 | No | Yes |
| KNN_11 | 60.911 | 0.676±0.106 | 0.607±0.312 | 0.565±0.138 | No | Yes |
| L_SVM | 50.392 | 0.252±0.252 | 0.500±0.500 | 0.335±0.335 | No | Yes |
| RBF SVM | 75.039 | 0.750±0.006 | 0.750±0.006 | 0.750±0.000 | No | Yes |
| DT | 69.231 | 0.693±0.010 | 0.692±0.016 | 0.692±0.003 | No | Yes |
| NN | 75.51 | 0.755±0.002 | 0.755±0.002 | 0.755±0.002 | No | Yes |
| NB | 66.091 | 0.661±0.001 | 0.661±0.009 | 0.661±0.005 | No | Yes |
| KNN_3 | 59.812 | 0.604±0.022 | 0.597±0.126 | 0.591±0.053 | Yes | No |
| KNN_7 | 61.381 | 0.625±0.036 | 0.613±0.157 | 0.603±0.064 | Yes | No |
| KNN_11 | 62.794 | 0.649±0.054 | 0.626±0.193 | 0.613±0.077 | Yes | No |
| L_SVM | 50.392 | 0.252±0.252 | 0.500±0.500 | 0.335±0.335 | Yes | No |
| RBF SVM | 74.568 | 0.746±0.016 | 0.746±0.026 | 0.746±0.005 | Yes | No |
| DT | 60.283 | 0.627±0.058 | 0.604±0.212 | 0.585±0.086 | Yes | No |
| NN | 76.138 | 0.762±0.010 | 0.761±0.014 | 0.761±0.002 | Yes | No |
| NB | 72.841 | 0.729±0.010 | 0.728±0.029 | 0.728±0.009 | Yes | No |
| KNN_3 | 61.695 | 0.700±0.127 | 0.614±0.326 | 0.570±0.143 | No | No |
| KNN_7 | 58.556 | 0.667±0.115 | 0.583±0.355 | 0.524±0.171 | No | No |
| KNN_11 | 59.969 | 0.709±0.150 | 0.597±0.366 | 0.536±0.172 | No | No |
| L_SVM | 50.392 | 0.252±0.252 | 0.500±0.500 | 0.335±0.335 | No | No |
| RBF SVM | 78.493 | 0.788±0.026 | 0.785±0.050 | 0.784±0.012 | No | No |
| DT | 66.876 | 0.669±0.002 | 0.669±0.004 | 0.669±0.003 | No | No |
| NN | 79.121 | 0.792±0.012 | 0.791±0.025 | 0.791±0.007 | No | No |
| NB | 75.981 | 0.760±0.004 | 0.760±0.013 | 0.760±0.005 | No | No |

5. Conclusions

In this paper, we approached the tasks of fake news and aggressiveness identification for the 2020 MEX-A3T contest. Using machine learning algorithms, we generated classification models for these tasks using different combinations of preprocessing techniques and keyword extraction methods. Our best configurations for both of the tasks are NN and RBF (SVM) with the TF-IDF method and without using the preprocessing techniques of removing the stop words and the stemming. As future work, we look forward to exploring other preprocessing techniques and keyword extraction methods to improve our ranking for the next MEX-AT3 contests.

Table 2

Performance results of the proposed models for the FakeNews dataset by using TF-IDF in the validation stage.

| Classifier | Accuracy | Precision | Recall | F-measure | Stopwords | Stemming |
|------------|---------------|--------------------|--------------------|--------------------|-----------|----------|
| KNN_3 | 57.614 | 0.585±0.025 | 0.575±0.170 | 0.563±0.076 | Yes | Yes |
| KNN_7 | 60.597 | 0.631±0.055 | 0.604±0.224 | 0.584±0.095 | Yes | Yes |
| KNN_11 | 62.009 | 0.651±0.067 | 0.618±0.232 | 0.597±0.095 | Yes | Yes |
| L_SVM | 50.392 | 0.252±0.252 | 0.500±0.500 | 0.335±0.335 | Yes | Yes |
| RBF SVM | 76.138 | 0.762±0.013 | 0.762±0.020 | 0.761±0.003 | Yes | Yes |
| DT | 65.62 | 0.658±0.014 | 0.656±0.055 | 0.655±0.021 | Yes | Yes |
| NN | 77.237 | 0.772±0.004 | 0.772±0.003 | 0.772±0.001 | Yes | Yes |
| NB | 65.777 | 0.658±0.002 | 0.658±0.006 | 0.658±0.004 | Yes | Yes |
| KNN_3 | 60.597 | 0.626±0.048 | 0.604±0.206 | 0.588±0.087 | No | Yes |
| KNN_7 | 60.597 | 0.639±0.066 | 0.604±0.250 | 0.579±0.107 | No | Yes |
| KNN_11 | 62.951 | 0.672±0.084 | 0.628±0.254 | 0.603±0.103 | No | Yes |
| L_SVM | 50.392 | 0.252±0.252 | 0.500±0.500 | 0.335±0.335 | No | Yes |
| RBF SVM | 76.138 | 0.762±0.010 | 0.761±0.014 | 0.761±0.002 | No | Yes |
| DT | 63.265 | 0.633±0.001 | 0.633±0.019 | 0.632±0.009 | No | Yes |
| NN | 78.022 | 0.780±0.003 | 0.780±0.001 | 0.780±0.001 | No | Yes |
| NB | 66.719 | 0.667±0.001 | 0.667±0.009 | 0.667±0.005 | No | Yes |
| KNN_3 | 62.951 | 0.633±0.021 | 0.629±0.091 | 0.626±0.036 | Yes | No |
| KNN_7 | 63.108 | 0.635±0.020 | 0.630±0.089 | 0.628±0.035 | Yes | No |
| KNN_11 | 63.579 | 0.645±0.036 | 0.635±0.135 | 0.629±0.052 | Yes | No |
| L_SVM | 50.392 | 0.252±0.252 | 0.500±0.500 | 0.335±0.335 | Yes | No |
| RBF SVM | 74.882 | 0.751±0.024 | 0.749±0.042 | 0.748±0.009 | Yes | No |
| DT | 63.736 | 0.663±0.067 | 0.639±0.193 | 0.624±0.071 | Yes | No |
| NN | 76.609 | 0.766±0.006 | 0.766±0.006 | 0.766±0.000 | Yes | No |
| NB | 73.155 | 0.732±0.007 | 0.731±0.023 | 0.731±0.008 | Yes | No |
| KNN_3 | 58.713 | 0.663±0.110 | 0.584±0.347 | 0.529±0.166 | No | No |
| KNN_7 | 55.102 | 0.640±0.110 | 0.548±0.405 | 0.460±0.221 | No | No |
| KNN_11 | 54.474 | 0.668±0.142 | 0.541±0.434 | 0.437±0.247 | No | No |
| L_SVM | 50.392 | 0.252±0.252 | 0.500±0.500 | 0.335±0.335 | No | No |
| RBF SVM | 78.65 | 0.787±0.011 | 0.787±0.014 | 0.786±0.002 | No | No |
| DT | 67.19 | 0.672±0.006 | 0.672±0.008 | 0.672±0.001 | No | No |
| NN | 81.476 | 0.815±0.008 | 0.815±0.017 | 0.815±0.004 | No | No |
| NB | 74.568 | 0.746±0.005 | 0.746±0.004 | 0.746±0.000 | No | No |

Acknowledgments

S. Arce-Cardenas gratefully acknowledges the financial support from Tecnológico Nacional de México (TecNM) under the project 9518.20-P (2rn3nx).

References

- [1] M. B. Yassein, S. Aljawarneh, Y. A. Wahsheh, Survey of online social networks threats and solutions, in: 2019 IEEE Jordan International Joint Conference on Electrical Engineering

- and Information Technology (JEEIT), 2019, pp. 375–380.
- [2] D. Theocharis, A. Bekiari, et al., Applying social network indicators in the analysis of verbal aggressiveness at the school, *Journal of Computer and Communications* 5 (2017) 169. doi:10.4236/jcc.2017.57015.
 - [3] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, Y. Chang, Abusive language detection in online user content, in: *Proceedings of the 25th International Conference on World Wide Web, WWW '16, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 2016*, p. 145–153. URL: <https://doi.org/10.1145/2872427.2883062>. doi:10.1145/2872427.2883062.
 - [4] A. Bovet, H. A. Makse, Influence of fake news in twitter during the 2016 us presidential election, *Nature Communications* 10 (2019) 7. doi:10.1038/s41467-018-07761-2.
 - [5] M. E. Aragón, H. Jarquín, M. Montes-y Gómez, H. J. Escalante, L. Villaseñor-Pineda, H. Gómez-Adorno, G. Bel-Enguix, J.-P. Posadas-Durán, Overview of mex-a3t at iberlef 2020: Fake news and aggressiveness analysis in mexican spanish, in: *Notebook Papers of 2nd SEPLN Workshop on Iberian Languages Evaluation Forum (IberLEF), Malaga, Spain, September, 2020*.
 - [6] J.-P. Posadas-Durán, H. Gómez-Adorno, G. Sidorov, J. J. M. Escobar, Detection of fake news in a new corpus for the spanish language, *Journal of Intelligent & Fuzzy Systems* 36 (2019) 4869–4876.
 - [7] M. Á. Álvarez-Carmona, E. Guzmán-Falcón, M. Montes-y Gómez, H. J. Escalante, L. Villaseñor-Pineda, V. Reyes-Meza, A. Rico-Sulayes, Overview of mex-a3t at ibereval 2018: Authorship and aggressiveness analysis in mexican spanish tweets, in: *Notebook Papers of 3rd SEPLN Workshop on Evaluation of Human Language Technologies for Iberian Languages (IBEREVAL), Seville, Spain, volume 6, 2018*.
 - [8] M. Graff, S. Miranda-Jiménez, E. S. Tellez, D. Moctezuma, V. Salgado, J. Ortiz-Bejar, C. N. Sánchez, Ingeotec at mex-a3t: Author profiling and aggressiveness analysis in twitter using μ tc and evomsa., in: *IberEval@ SEPLN, 2018*, pp. 128–133.
 - [9] S. Correa, A. Martin, Linguistic generalization of slang used in mexican tweets, applied in aggressiveness detection., in: *IberEval@ SEPLN, 2018*, pp. 119–127.
 - [10] M. E. Aragón, M. Á. Álvarez-Carmona, M. Montes-y Gómez, H. J. Escalante, L. Villaseñor-Pineda, D. Moctezuma, Overview of mex-a3t at iberlef 2019: Authorship and aggressiveness analysis in mexican spanish tweets, in: *Notebook Papers of 1st SEPLN Workshop on Iberian Languages Evaluation Forum (IberLEF), Bilbao, Spain, 2019*.
 - [11] M. Casavantes, R. López, L. C. González, Uach at mex-a3t 2019: Preliminary results on detecting aggressive tweets by adding author information via an unsupervised strategy, in: *In Proceedings of the First Workshop for Iberian Languages Evaluation Forum (IberLEF 2019), CEUR WS Proceedings, 2019*.
 - [12] H. Q. Abonizio, J. I. de Moraes, G. M. Tavares, S. Barbon Junior, Language-independent fake news detection: English, portuguese, and spanish mutual features, *Future Internet* 12 (2020) 87.

Table 3

Performance results of the proposed models for the Aggressiveness dataset by using TF in the validation stage.

| Classifier | Accuracy | Precision | Recall | F-measure | Stopwords | Stemming |
|------------|---------------|--------------------|--------------------|--------------------|-----------|----------|
| KNN_3 | 75.762 | 0.713±0.067 | 0.641±0.278 | 0.654±0.190 | Yes | Yes |
| KNN_7 | 75.99 | 0.738±0.028 | 0.621±0.331 | 0.631±0.218 | Yes | Yes |
| KNN_11 | 75.686 | 0.741±0.020 | 0.611±0.348 | 0.617±0.232 | Yes | Yes |
| L_SVM | 72.031 | 0.806±0.088 | 0.519±0.479 | 0.456±0.380 | Yes | Yes |
| RBF SVM | 81.344 | 0.811±0.004 | 0.711±0.243 | 0.736±0.143 | Yes | Yes |
| DT | 78.735 | 0.768±0.029 | 0.676±0.264 | 0.696±0.167 | Yes | Yes |
| NN | 81.435 | 0.805±0.015 | 0.718±0.228 | 0.742±0.137 | Yes | Yes |
| NB | 60.413 | 0.641±0.232 | 0.667±0.149 | 0.597±0.053 | Yes | Yes |
| KNN_3 | 75.171 | 0.704±0.069 | 0.629±0.292 | 0.640±0.200 | No | Yes |
| KNN_7 | 75.034 | 0.718±0.042 | 0.607±0.341 | 0.613±0.231 | No | Yes |
| KNN_11 | 75.049 | 0.727±0.030 | 0.601±0.355 | 0.604±0.241 | No | Yes |
| L_SVM | 71.53 | 0.809±0.095 | 0.510±0.490 | 0.436±0.397 | No | Yes |
| RBF SVM | 81.556 | 0.817±0.002 | 0.713±0.245 | 0.738±0.143 | No | Yes |
| DT | 78.553 | 0.767±0.028 | 0.672±0.270 | 0.691±0.170 | No | Yes |
| NN | 81.283 | 0.808±0.008 | 0.712±0.240 | 0.736±0.142 | No | Yes |
| NB | 62.096 | 0.648±0.228 | 0.677±0.134 | 0.612±0.058 | No | Yes |
| KNN_3 | 77.825 | 0.735±0.075 | 0.691±0.208 | 0.705±0.147 | Yes | No |
| KNN_7 | 77.886 | 0.744±0.055 | 0.676±0.244 | 0.693±0.162 | Yes | No |
| KNN_11 | 78.083 | 0.761±0.028 | 0.663±0.279 | 0.682±0.178 | Yes | No |
| L_SVM | 73.199 | 0.807±0.080 | 0.541±0.455 | 0.499±0.342 | Yes | No |
| RBF SVM | 81.116 | 0.796±0.025 | 0.718±0.222 | 0.740±0.136 | Yes | No |
| DT | 77.158 | 0.775±0.005 | 0.631±0.335 | 0.643±0.215 | Yes | No |
| NN | 80.965 | 0.797±0.019 | 0.712±0.232 | 0.735±0.141 | Yes | No |
| NB | 63.325 | 0.650±0.221 | 0.681±0.113 | 0.622±0.065 | Yes | No |
| KNN_3 | 74.867 | 0.692±0.091 | 0.644±0.250 | 0.655±0.179 | No | No |
| KNN_7 | 76.399 | 0.727±0.054 | 0.645±0.283 | 0.659±0.189 | No | No |
| KNN_11 | 76.824 | 0.741±0.037 | 0.643±0.297 | 0.658±0.194 | No | No |
| L_SVM | 71.045 | 0.730±0.020 | 0.501±0.499 | 0.417±0.414 | No | No |
| RBF SVM | 81.753 | 0.817±0.001 | 0.717±0.239 | 0.742±0.139 | No | No |
| DT | 77.977 | 0.785±0.007 | 0.645±0.319 | 0.662±0.200 | No | No |
| NN | 81.435 | 0.809±0.008 | 0.715±0.236 | 0.739±0.140 | No | No |
| NB | 64.22 | 0.656±0.220 | 0.689±0.112 | 0.631±0.066 | No | No |

Table 4

Performance results of the proposed models for the Aggressiveness dataset by using TF-IDF in the validation stage.

| Classifier | Accuracy | Precision | Recall | F-measure | Stopwords | Stemming |
|------------|---------------|--------------------|--------------------|--------------------|-----------|----------|
| KNN_3 | 74.063 | 0.692±0.065 | 0.598±0.339 | 0.602±0.235 | Yes | Yes |
| KNN_7 | 74.215 | 0.725±0.021 | 0.579±0.388 | 0.571±0.271 | Yes | Yes |
| KNN_11 | 73.639 | 0.728±0.010 | 0.563±0.413 | 0.544±0.297 | Yes | Yes |
| L_SVM | 71.728 | 0.838±0.123 | 0.513±0.487 | 0.442±0.392 | Yes | Yes |
| RBF SVM | 80.98 | 0.813±0.004 | 0.701±0.258 | 0.726±0.151 | Yes | Yes |
| DT | 78.538 | 0.765±0.030 | 0.673±0.267 | 0.692±0.169 | Yes | Yes |
| NN | 81.283 | 0.806±0.011 | 0.714±0.236 | 0.737±0.141 | Yes | Yes |
| NB | 60.458 | 0.639±0.231 | 0.665±0.144 | 0.597±0.055 | Yes | Yes |
| KNN_3 | 73.881 | 0.701±0.047 | 0.582±0.372 | 0.578±0.260 | No | Yes |
| KNN_7 | 73.487 | 0.736±0.001 | 0.557±0.424 | 0.532±0.308 | No | Yes |
| KNN_11 | 73.047 | 0.726±0.005 | 0.548±0.434 | 0.518±0.321 | No | Yes |
| L_SVM | 71.455 | 0.826±0.113 | 0.508±0.492 | 0.432±0.400 | No | Yes |
| RBF SVM | 81.42 | 0.821±0.010 | 0.707±0.256 | 0.732±0.148 | No | Yes |
| DT | 78.492 | 0.766±0.028 | 0.671±0.270 | 0.690±0.171 | No | Yes |
| NN | 81.541 | 0.806±0.015 | 0.720±0.227 | 0.743±0.136 | No | Yes |
| NB | 62.051 | 0.650±0.229 | 0.679±0.139 | 0.612±0.057 | No | Yes |
| KNN_3 | 76.596 | 0.719±0.078 | 0.669±0.232 | 0.683±0.163 | Yes | No |
| KNN_7 | 77.233 | 0.745±0.039 | 0.652±0.285 | 0.669±0.185 | Yes | No |
| KNN_11 | 77.582 | 0.776±0.000 | 0.641±0.322 | 0.655±0.204 | Yes | No |
| L_SVM | 73.093 | 0.823±0.097 | 0.538±0.459 | 0.493±0.348 | Yes | No |
| RBF SVM | 81.132 | 0.798±0.021 | 0.716±0.227 | 0.738±0.138 | Yes | No |
| DT | 76.885 | 0.772±0.004 | 0.626±0.341 | 0.636±0.219 | Yes | No |
| NN | 81.04 | 0.797±0.021 | 0.715±0.228 | 0.737±0.139 | Yes | No |
| NB | 63.598 | 0.648±0.218 | 0.679±0.103 | 0.623±0.069 | Yes | No |
| KNN_3 | 76.96 | 0.741±0.041 | 0.648±0.288 | 0.664±0.189 | No | No |
| KNN_7 | 78.508 | 0.775±0.014 | 0.664±0.288 | 0.683±0.180 | No | No |
| KNN_11 | 79.478 | 0.797±0.003 | 0.675±0.286 | 0.696±0.173 | No | No |
| L_SVM | 71.318 | 0.833±0.121 | 0.505±0.494 | 0.427±0.405 | No | No |
| RBF SVM | 82.345 | 0.827±0.006 | 0.725±0.235 | 0.751±0.134 | No | No |
| DT | 78.159 | 0.790±0.011 | 0.647±0.319 | 0.664±0.199 | No | No |
| NN | 82.345 | 0.820±0.005 | 0.730±0.223 | 0.754±0.130 | No | No |
| NB | 65.418 | 0.659±0.214 | 0.693±0.094 | 0.640±0.071 | No | No |