


Classification of Public Administration Complaints

Francisco Caldeira ✉

Iscte, University Institute of Lisbon, Portugal

Luís Nunes ✉ 

Iscte, University Institute of Lisbon, Portugal

ISTAR, Lisbon, Portugal

Ricardo Ribeiro ✉ 

Iscte, University Institute of Lisbon, Portugal

INESC-ID Lisbon, Portugal

Abstract

Complaint management is a problem faced by many organizations that is both vital to customer image and highly dependent on human resources. This work attempts to tackle a part of the problem, by classifying summaries of complaints using machine learning models in order to better redirect these to the appropriate responders. The main challenges of this task is that training datasets are often small and highly imbalanced. This can have a big impact on the performance of classification models. The dataset analyzed in this work suffers from both of these problems, being relatively small and having labels in different proportions. In this work, two different techniques are analyzed: combining classes together to increase the number of elements of the new class; and, providing new artificial examples for some classes via translation into other languages. The classification models explored were the following: k -NN, SVM, Naïve Bayes, boosting, and Deep Learning approaches, including transformers. The paper concludes that although, as expected, the classes with little representation are hard to classify, the techniques explored helped to boost the performance, especially in the classes with a low number of elements. SVM and BERT-based models outperformed their peers.

2012 ACM Subject Classification Information systems → Clustering and classification

Keywords and phrases Text Classification, Natural Language Processing, Deep Learning, BERT

Digital Object Identifier 10.4230/OASICS.SLATE.2022.9

Funding This work was partly funded through national funds by FCT - Fundação para a Ciência e Tecnologia, I.P. under project UIDB/04466/2020 (ISTAR).

1 Introduction

The process of manual classification of user generated data can be regarded as a tedious and error prone task. In the attempt to render this process more efficiently, this paper explores the use of Machine Learning and Natural Language Processing tools and models for this task.

This work will focus on the classification of summarized complaints received by a Portuguese public institute. After being received, the complaints are read and manually processed by a worker from the public institute. Then, after being correctly classified, the complaint is forwarded to the respective department for additional processing, accompanied by a textual summary. We concentrate on the classification of summaries for redirection to the appropriate responders. The final goal is to render assistance in the whole process. Due to privacy questions, the dataset is not available.

The related work is presented in Section 2. Data is explored in Section 3, along with the techniques used for processing, providing some insights on the structure and contents of the dataset. The implementation is described in Section 4. In Section 5 the achieved results are presented. Finally, the document closes with the conclusions and some possible future work in Section 6.



© Francisco Caldeira, Luís Nunes, and Ricardo Ribeiro;
licensed under Creative Commons License CC-BY 4.0

11th Symposium on Languages, Applications and Technologies (SLATE 2022).

Editors: João Cordeiro, Maria João Pereira, Nuno F. Rodrigues, and Sebastião Pais; Article No. 9; pp. 9:1–9:12

OpenAccess Series in Informatics



OASICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

2 Related Work

Complaint data can be viewed and analyzed as a “form of user generated data” and thus the current state of the art for processing user commentaries of social media content can be deemed useful for the task analyzed in this work.

A similar problem to ours was also tackled by Lopes-Cardoso et al. [5]. This work focus on the classification of complaint data from another Portuguese public institute in three different dimensions: economic activity prediction, infraction severity prediction, and competence prediction. The authors reported SVM with a linear kernel was the best performing model with around 79% accuracy on the competence prediction task, which is the one most similar to ours. In a comparable stream of work [7, 8, 2], the focus was also the classification of data from a multitude of Portuguese public services while also dealing with noisy and imbalanced data. The best results were achieved by SVM and BERT-based approaches. Oliveira et al. [4] used a set of frequently asked questions from a Portuguese public institute for a classification task. To create additional data the authors used Google API to translate back and forth to Portuguese and English to develop new data. They also used native speakers to create and improve the alternative constructions of the questions. In their results, the best performing model was SVM with a better training time/performance ratio when compared to other models. In the end they also noted that using a fine tuned BERT model improved the results at cost of higher train time. For benchmark they used the set of the generated question to match the original ones and a classification based of the area of business of each question. In [3] the authors performed an analysis of user commentaries from a Portuguese telecommunication company for a sentiment analysis task. They noted that in stemming in Portuguese is not very useful and resorted to custom built list of rules to work along side the stemming to reduce edge cases. In [6], the authors compared SVM with Universal Language Model Fine-tuning (ULMFiT), for classification of official Brazilian Government data. The concluded that, even though ULMFiT is a state-of-the-art technique for classification, it only corresponded to a small increase in classification accuracy when compared to the SVM model.

Wang et al. [13] explore “*Label-Embedding Attentive Model*” (LEAM) by proposing a word embedding approach in which both words and labels are joined in the same latent space in order to measure the compatibility of word-label used as document representations. While in some cases the LEAM model was unable to out perform other models the authors stated the algorithm is much less demanding in comparison to other state-of-the-art algorithms. Tang et al. [12] also worked in classification of complaints and proposed a combination of BERT and word2vec models to try and improve the overall accuracy. To tackle imbalanced data they experimented with translations of the text to expand the training data.

Further expanding on deep learning strategies, a specific BERT model was trained on Brazilian Portuguese data, the model was named BERTimbau [11]. When comparing BERTimbau performance against BERT for Named Entity Recognition tasks and Sentence Textual Similarity the first would outperform the latter with the authors stating “*large pre-trained learning models can be valuable assets especially for languages that have few annotated resources but abundant unlabeled data, such as Portuguese*” [11].

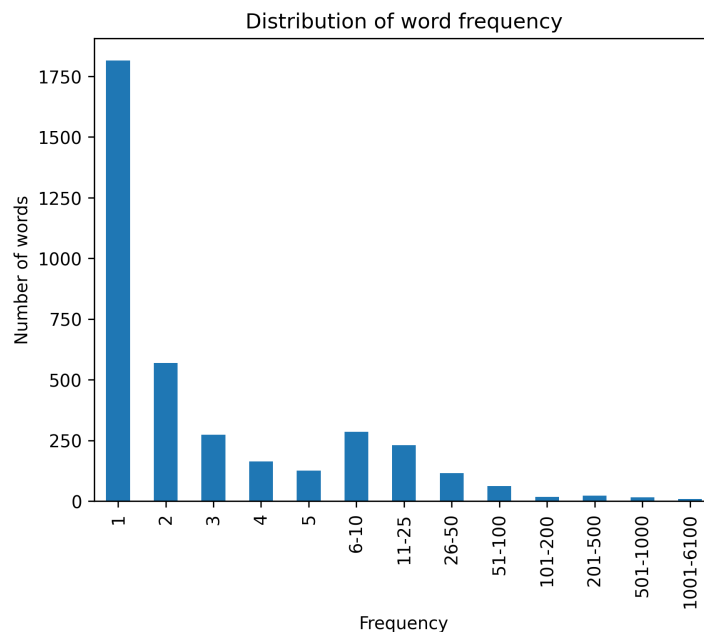
Also worth considering a portuguese wordnet, semantically structured lexical database, expansion done in [10]. The word database is useful for extracting synonyms to create more artificial entries in the data set.

For this work, the techniques used to enhance are tested in combination with the Portuguese text processing strategies in order to tackle the classification task. With the main goal of understanding the benefit of these techniques.

3 Data

The complaints dataset features 4459 complaints gathered from 2020 to 2021 and spread over 17 different labels. Each category directly references to an entity governed by the portuguese Ministry of Justice and are listed here <https://igsj.justica.gov.pt/Servicos/Apresentar-queixa>. Not all entities are featured in the dataset since not all receive complaints. The complaints were subject to an initial processing from the institute that summarized the content of the complaint into a short text, leaving mostly the relevant words in the text. The dataset featured two main columns used for the classification task, the complaint and the assigned department.

The summarized complaints have an average of 16 words per complaint with the shortest having only 2 words and longest featuring 38 words, all of them written in Portuguese. The smallest summaries (of only two words) was repeated 2 times and the text can be seen as *"process delays"* and the summaries of with 3 and 4 words were similar with the added words giving more detail regarding the actual complaint. Of the 4459 summarized complaints it was extracted a corpus of 3705 different tokens, the full distribution can be seen in Figure 1, with 1805 tokens only appearing once and two appearing almost all summaries.



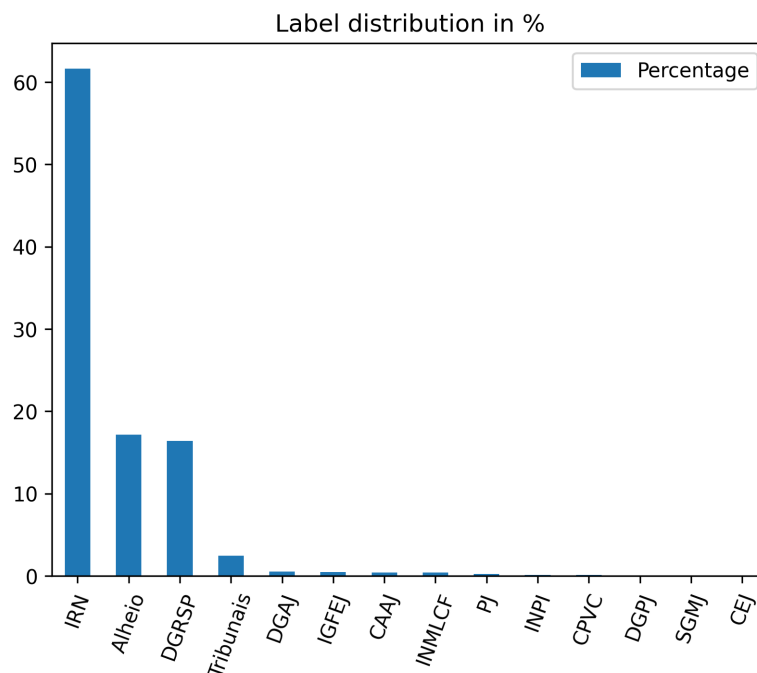
■ **Figure 1** Word frequency distribution.

The dataset presents high variability in what regards label cardinality: the category with more complaints covers almost 60% of the dataset while the second and third classes only account for 35%. Six labels also contained ten or less complaints with a single class only containing one complaint.

One single complaint was assigned two different classes and was re-assigned to the class that presented the lowest number of complaints.

Since the text data was preprocessed by a worker of the public institute, the text data features little spelling errors unlike some other user generated data like social media content.

9:4 Classification of Public Administration Complaints



■ **Figure 2** Label distribution of the data explored, more than 90% of the data was labeled to the top 3 labels.

4 Experiments

In this section the text processing tasks and the techniques used to handle the data imbalance are detailed.

Considering the properties of the dataset, both traditional classification and deep learning approaches are compared. We aim to better understand the performance of the experimented models with the techniques used for dealing with imbalanced data.

4.1 Data Preparation

Handling data imbalance was achieved with two different strategies, translating the texts of the complaints and creating additional labels to group classes with lower cardinality.

To increase the number of class representatives, the documents of the classes that had between 10 and 30 elements were translated into several languages (English, Spanish, Italian, Polish, and German) and back to Portuguese. This strategy increased the number of representatives by sixfold (Table 1). The artificially produced complaints were only used for model training to ensure the validation was performed with real world data.

For the classification experiments both datasets needed to undergo a battery of processing steps in order to be used as input for the models:

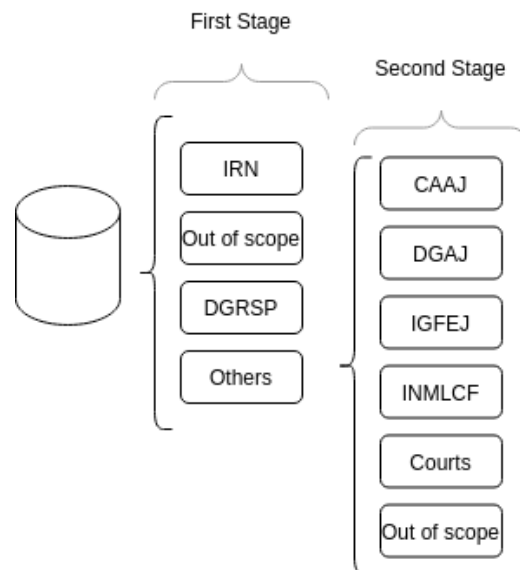
- lower casing all characters;
- removing numerical data;
- removing special characters;
- removing diacritics.

As noted in [3] stemming is not very useful for the portuguese language was not considered for this task.

The texts also featured some acronyms that could be more insightfully expanded. For example, “cc” was replaced to match “cartao do cidadao” (identity card) and “ep” was replaced by “estabelecimento prisional” (prison establishment).

The documents were then regrouped into different labels. Firstly, the complaints were separated into the top 3 classes and the remaining classes. For the final split, the top 4 to the top 8 classes were given the respective label and labels that had less than 10 example were combined into a single class. Figure 3 illustrates this process. From this point on wards, the first stage classification will refer the classification of the top 3 classes along with the “Others” (new label for the remaining classes) and the second stage classification will refer to classification of the others to the remaining considered classes. The full data classification will refer to the top 8 classes only along with the remaining grouped into the class “Others”. For all the experiments, the classes with a cardinality lower than 15 were not considered as a full class: they were assigned the label of “Others”.

The groups were assembled based on the labels distribution in Figure 2 with the first stage containing only the more frequent labels and for the second stage the split were the categories with more than 15 examples, the size of the classes can be seen in Table 2, the complete dataset without any segmentation was used to compared the performance of the multistage classification.



■ **Figure 3** To handle the label imbalance issue a multistage classification method was analyzed, the top 3 classes were initially classified and the remaining top 5 classes were also considered, class with lower representation were not considered for the task.

The translation technique was particularly useful. When translating back to Portuguese some synonyms appeared adding more words to the corpus and complementing the shortcomings of the less populated labels. After increasing the number of examples, the dataset featured 5000 complaints, almost an increase of 500 new complaints and an extra 300 new tokens.

■ **Table 1** Distribution of complaints per label, the first 3 labels contain almost 80% of the data set. Using the translation technique made the classes more equal in regards to their size.

	Class	Class with translations
IRN	2748	2748
Alheio	765	765
DGRSP	732	732
Courts	109	109
DGAJ	22	132
IGFEJ	21	123
CAAJ	19	114
INMLCF	17	102
PJ	10	60
INPI	5	30
CPVC	5	30
DGPJ	3	18
SGMJ	2	12
CEJ	1	6

4.2 Experimental Setup

The complaints were processed using the standard tokenization pipelines (special characters were removed). All characters were lowercased and the TF-IDF model was used to get the feature vectors for each complaint. For the BERT-based approach, instead of TD-IDF the tokens were preprocessed using the model encodings. For additional testing and comparison the words that only appeared once were removed as well as the top 3 words.

The experimented models were evaluated using the standard metrics with the goal to compare performance gain by using the techniques referred to handle the highlighted issues: low cardinality of the dataset and the class imbalance that was previously mentioned.

From here the experiments were separated into two distinct experiences, having the translations in the training and the original complaints in order to evaluate the performance gain by using the fabricated examples.

The dataset was split into two groups 30% for testing, 70% was used for training and from the training set 15% was used for validation. For the deep learning models, the training set was split into an additional validation set for training purposes. For comparison, the training set for the original dataset had 520 examples for the first stage while the expanded training set featured 1713. This difference is even noticeable in the second stage training from 56 to 169 examples for the training, refer to Table 2.

■ **Table 2** Number of inputs for the different tasks for each model.

	Classification	Original	Augmented
Training	Full	84	627
	First Stage	520	1713
	Second Stage	56	169
Validation	Full	15	111
	First Stage	92	303
	Second Stage	10	74
Testing		1338	

4.3 Methods

As previously mentioned, given the properties of the dataset, we experimented several classification models, ranging from more traditional approaches like Naïves Bayes, k -NN, and SVM models to deep learning-based approaches as Multilingual-BERT and XGBoosting.

To validate and optimize the hyper parameters of the models it was used cross-validation, splitting the data into multiple smaller subsets with equal class cardinality to validate the classifier while also avoiding to overfit the experimented models.

Several classification pipelines were also tested, classifying the full data set, only the first stage classification, and performing a multistage classification as illustrated in Figure 3.

5 Results and Discussion

The results for the classification and using the original dataset can be viewed in the Tables 3, 4, and 5.

■ **Table 3** Results for the first stage using the original dataset, SVM outperform all the other models for this task but Naïve Bayes presented a marginally higher f-score.

Model	Processing	Accuracy	Precision	Recall	F-score
Naive Bayes	No special processing	0.894619	0.921471	0.894619	0.903991
	Removing stopwords, low frequency words	0.902093	0.930147	0.902093	0.911670
SVM	No special processing	0.899851	0.929783	0.899851	0.910339
	Removing stopwords, low frequency words	0.894619	0.934678	0.894619	0.908281
k -NN	No special processing	0.853513	0.887257	0.853513	0.867000
	Removing stopwords, low frequency words	0.855007	0.895267	0.855007	0.870560
XGBoost	No special processing	0.826607	0.883714	0.826607	0.848738
	Removing stopwords, low frequency words	0.828102	0.903759	0.828102	0.856595
Multilingual-Bert	No special processing	0.869207	0.91050	0.86920	0.883518
	Removing stopwords, low frequency words	0.857249	0.9442	0.857249	0.88504

Considering the first stage classification for the original dataset, the deep learning models and the more traditional models yielded similar results to SVM, proving to be the model with best performance in this experiment.

For the second stage classification, the results were the worst from all the experiments. While presenting around 70% precision, the accuracy and f-score values were extremely low. Due to low number of class representatives, the BERT model was unable to be fine tuned. The training and validation data would be very short for the complexity of a neural network.

For the full classification when using the original dataset, k -NN and SVM presented good results with an 80% accuracy and 92% precision, respectively. When comparing the results for each class in Table 6 it becomes apparent the good results are heavily weighted by the distribution of the testing set. The classes with more representatives have more weight and yield better performance.

The deep learning models had an accuracy value of 58% when compared to k -NN with 80% accuracy.

■ **Table 4** Results for the second stage classification using the original dataset, low performance across all models. BERT model was unable to trained for this stage.

Model	Processing	Accuracy	Precision	Recall	F-score
SVM	No special processing	0.312977	0.672483	0.312977	0.246029
	Removing stopwords, low frequency words	0.293333	0.700265	0.293333	0.242265
Naive Bayes	No special processing	0.259843	0.577479	0.259843	0.181176
	Removing stopwords, low frequency words	0.271318	0.753100	0.271318	0.188976
XGBoost	No special processing	0.166667	0.606909	0.166667	0.184742
	Removing stopwords, low frequency words	0.304569	0.791608	0.304569	0.370414
<i>k</i> -NN	No special processing	0.230769	0.738754	0.230769	0.194529
	Removing stopwords, low frequency words	0.312500	0.809010	0.312500	0.311697

Considering only the first stage classification task and not using, the results were much better results with 89% accuracy using SVM, although only for a limited number of classes. The full classification, while also presenting good results, was skewed from the imbalance of the testing set, most of the accuracy was attributed to three labels.

The results for the classification using the expanded dataset can be viewed in Tables 7, 8, and 9. The expanded dataset yielded considerably better results for all the tasks examined in this work.

The first stage classification had an increase of almost 6% from the original data set with SVM outperforming the others. Surprisingly the second stage classification almost doubling the performance achieved with the original dataset. Accuracy improved to 58% from 29% and f-score increased to 52% from 24%, when using an SVM. With the augmented dataset it was possible to fine-tune BERT for the second classification and it proved to be the best performing model for this task.

For the full classification using SVM, the accuracy reported was of 89% with 90% f-score, an increase of 10p.p. from the dataset. Using BERT for this task returned similar results to the SVM noting the approximated 20p.p. increase in accuracy from the original dataset, from 58% to 76%. Having a bigger training pool was essential in improving the BERT model.

6 Conclusion and Future Work

Similar to the results achieved by Silva et al. [9], the concise versions of the complaints proved to be enough for a reasonable and usable classification result, in the data explored removing stopwords and frequent words had low gains. In this case, since the tokens were already part of a short corpus, the additional processing (removing stopwords, the most highest and low frequency words) showed only marginal improvements. The classes with more representatives were the ones that achieved a higher score in precision, as the complaints were more similar and mostly relating to a more frequent issue. The categories with lower representatives were generally harder to classify as the texts and issues featured little resemblance.

Due to the low number of complaints for some classes, the more traditional models had a better performance than the deep learning models, especially when using the original dataset. When using the expanded dataset, deep learning models yielded similar results to more traditional models. For the second stage classification BERT was able to outperform

■ **Table 5** Results for the full classification using the original dataset, k -NN and SVM have similar scores and present decent performance although SVM has a higher precision and f-score.

Model	Processing	Accuracy	Precision	Recall	F-score
Naive Bayes	No special processing	0.750374	0.907025	0.750374	0.810925
	Removing stopwords, low frequency words	0.770553	0.92491	0.770553	0.832805
SVM	No special processing	0.786996	0.913812	0.786996	0.837782
	Removing stopwords, low frequency words	0.793722	0.924617	0.793722	0.845285
KNN	No special processing	0.796712	0.887542	0.796712	0.833793
	Removing stopwords, low frequency words	0.803438	0.883222	0.803438	0.837429
XGBoost	No special processing	0.601644	0.842656	0.601644	0.685857
	Removing stopwords, low frequency words	0.633782	0.894219	0.633782	0.730709
Multilingual-Bert	No special processing	0.58071	0.82905	0.58071	0.67277
	Removing stopwords, low frequency words	0.550822	0.84441	0.550822	0.65165

■ **Table 6** Metrics by class for SVM using the original dataset.

	Precision	Recall	F-Score	Support
Alheio	0.88	0.64	0.74	247
CAAJ	0.14	0.83	0.24	6
DGAJ	0.05	0.43	0.10	7
DGRSP	0.97	0.90	0.94	228
IGFEJ	0.11	0.75	0.19	8
INMLCF	0.13	0.83	0.22	6
IRN	0.98	0.82	0.89	802
Others	0.11	0.57	0.18	7
Courts	0.27	0.52	0.35	27

all of the others. The expanded dataset led to considerably better results when compared to the original data set, especially for the second stage classification task and for the full classification. Boosting the number of representatives for each class, especially the least represented ones, greatly improved the performance of the models. The new sentences and the tokens introduced were essential to improve the BERT performance. Even though further testing is needed to confirm this, these experiments seem to indicate that balancing the data played a crucial role in the performance gains and could be considered as technique for improving datasets with lower cardinality.

For future work, the fine tuning of a multistage classification method should be explored while also considering more classification stages and using binary classification [1]. More techniques for expanding the corpus and classes examples could also be explored, as an example wordnets could be used to further diversify the dataset. Producing hand made examples for a class could also provide better representatives for each class as the machine made translations that can only reach a certain limit. It should also be noted that combining different models for the various stages of the classification could provide additional insights.

9:10 Classification of Public Administration Complaints

■ **Table 7** Results for the first stage classification using the augmented dataset, SVM outperform all the other models for this task.

Model	Processing	Accuracy	Precision	Recall	F-score
Naive Bayes	No special processing	0.930493	0.932668	0.930493	0.931215
	Removing stopwords, low frequency words	0.933483	0.937128	0.933483	0.934630
SVM	No special processing	0.934230	0.941662	0.934230	0.937099
	Removing stopwords, low frequency words	0.940957	0.945378	0.940957	0.942697
<i>k</i> -NN	No special processing	0.911809	0.914773	0.911809	0.911943
	Removing stopwords, low frequency words	0.904335	0.908641	0.904335	0.905653
XGBoost	No special processing	0.904335	0.918122	0.904335	0.909718
	Removing stopwords, low frequency words	0.912556	0.924713	0.912556	0.917385
Multilingual-Bert	No special processing	0.93	0.93025	0.93	0.92831
	Removing stopwords, low frequency words	0.936472	0.93789	0.936472	0.93710

■ **Table 8** Results for the second stage classification using the expanded dataset, BERT outperform all the other models for this task.

Model	Processing	Accuracy	Precision	Recall	F-score
SVM	Removing stopwords, low frequency words	0.578947	0.767832	0.578947	0.525620
	No special processing	0.488095	0.772963	0.488095	0.405475
Naive Bayes	Removing stopwords, low frequency words	0.447368	0.480623	0.447368	0.335068
	No special processing	0.472222	0.432330	0.472222	0.370383
XGBoost	Removing stopwords, low frequency words	0.534884	0.716058	0.534884	0.535191
	No special processing	0.455556	0.718620	0.455556	0.447290
<i>k</i> -NN	Removing stopwords, low frequency words	0.487500	0.751803	0.487500	0.405868
	No special processing	0.432432	0.775594	0.432432	0.330564
Multilingual-Bert	No special processing	0.65217	0.64684	0.65217	0.59163
	Removing stopwords, low frequency words	0.55384	0.7719	0.55384	0.483812

■ **Table 9** Results for the full classification using the expanded dataset, SVM outperform all the other models for this task.

Model	Processing	Accuracy	Precision	Recall	F-score
Naive Bayes	No special processing	0.870703	0.931320	0.870703	0.891887
	Removing stopwords, low frequency words	0.878924	0.931459	0.878924	0.896761
SVM	No special processing	0.884155	0.921298	0.884155	0.897685
	Removing stopwords, low frequency words	0.890882	0.924622	0.890882	0.902759
<i>k</i> -NN	No special processing	0.835575	0.908580	0.835575	0.860282
	Removing stopwords, low frequency words	0.837818	0.909102	0.837818	0.863215
XGBoost	No special processing	0.774290	0.877428	0.774290	0.814150
	Removing stopwords, low frequency words	0.791480	0.881701	0.791480	0.825520
Multilingual-Bert	No special processing	0.76233	0.9231	0.76233	0.819744
	Removing stopwords, low frequency words	0.84679	0.93258	0.84679	0.88095

References

- 1 Fernando Batista and Ricardo Ribeiro. Sentiment analysis and topic classification based on binary maximum entropy classifiers. *Proces. del Leng. Natural*, 50:77–84, 2013. URL: <http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/4662>.
- 2 André Fazendeiro. Automatic correspondence distribution for a public institution. Master’s thesis, Instituto Superior Técnico, 2021.
- 3 Ana Catarina Forte and Pavel B. Brazdil. Determining the level of clients’ dissatisfaction from their commentaries. In João Silva, Ricardo Ribeiro, Paulo Quaresma, André Adami, and António Branco, editors, *Computational Processing of the Portuguese Language*, pages 74–85, Cham, 2016. Springer International Publishing.
- 4 Hugo Gonçalo Oliveira, João Ferreira, José Santos, Pedro Fialho, Ricardo Rodrigues, Luisa Coheur, and Ana Alves. AIA-BDE: A corpus of FAQs in Portuguese and their variations. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5442–5449, Marseille, France, May 2020. European Language Resources Association. URL: <https://aclanthology.org/2020.lrec-1.669>.
- 5 Henrique Lopes-Cardoso, Tomás Freitas Osório, Luís Vilar Barbosa, Gil Rocha, Luís Paulo Reis, João Pedro Machado, and Ana Maria Oliveira. Robust complaint processing in portuguese. *Information*, 12(12), 2021. doi:10.3390/info12120525.
- 6 Pedro Henrique Luz de Araujo, Teófilo Emidio de Campos, and Marcelo Magalhães Silva de Sousa. Inferring the source of official texts: Can svm beat ulmfit? In Paulo Quaresma, Renata Vieira, Sandra Aluísio, Helena Moniz, Fernando Batista, and Teresa Gonçalves, editors, *Computational Processing of the Portuguese Language*, pages 76–86, Cham, 2020. Springer International Publishing.
- 7 Luis Neto. Cia: Citizen contact center agent assistant. Master’s thesis, Instituto Superior Técnico, January 2021.
- 8 Vilma Neves. Automatic classification of correspondence from a public institution. Master’s thesis, Instituto Superior Técnico, 2021.
- 9 Sara Silva, Ricardo Ribeiro, and Rúben Pereira. Less is more in incident categorization. In Pedro Rangel Henriques, José Paulo Leal, António Menezes Leitão, and Xavier Gómez Guinovart, editors, *7th Symposium on Languages, Applications and Technologies, SLATE 2018, June 21-22, 2018, Guimarães, Portugal*, volume 62 of *OASICs*, pages 17:1–17:7. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2018. doi:10.4230/OASICs.SLATE.2018.17.

9:12 Classification of Public Administration Complaints

- 10 Alberto Simões, Xavier Gómez Guinovart, and José João Almeida. Enriching a portuguese wordnet using synonyms from a monolingual dictionary. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, May 2016. European Language Resources Association (ELRA).
- 11 Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. Bertimbau: Pretrained bert models for brazilian portuguese. In Ricardo Cerri and Ronaldo C. Prati, editors, *Intelligent Systems*, pages 403–417, Cham, 2020. Springer International Publishing.
- 12 Xiaobo Tang, Hao Mou, Jiangnan Liu, and Xin Du. Research on automatic labeling of imbalanced texts of customer complaints based on text enhancement and layer-by-layer semantic matching. *Scientific Reports*, 11(1):11849, June 2021. doi:10.1038/s41598-021-91189-0.
- 13 Guoyin Wang, Chunyuan Li, Wenlin Wang, Yizhe Zhang, Dinghan Shen, Xinyuan Zhang, Ricardo Henao, and Lawrence Carin. Joint embedding of words and labels for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2321–2331, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi:10.18653/v1/P18-1216.