# Complaint Analysis and Classification for Economic and Food Safety

**João Filgueiras*, Luís Barbosa*, Gil Rocha*, Henrique Lopes Cardoso*, Luís Paulo Reis*,**
**João Pedro Machado+, Ana Maria Oliveira+**
*Laboratório de Inteligência Artificial e Ciência de Computadores (LIACC)
Faculdade de Engenharia da Universidade do Porto
Rua Dr. Roberto Frias, s/n, 4200-465 Porto, Portugal
+Autoridade de Segurança Alimentar e Económica (ASAE),
Rua Rodrigo da Fonseca, 73, 1269-274 Lisboa, Portugal
{filgueiras, up201405729, gil.rocha, hlc, lpreis}@fe.up.pt
{jpmachado, amoliveira}@asae.pt

## Abstract

Governmental institutions are employing artificial intelligence techniques to deal with their specific problems and exploit their huge amounts of both structured and unstructured information. In particular, natural language processing and machine learning techniques are being used to process citizen feedback. In this paper, we report on the use of such techniques for analyzing and classifying complaints, in the context of the Portuguese Economic and Food Safety Authority. Grounded in its operational process, we address three different classification problems: target economic activity, implied infraction severity level, and institutional competence. We show promising results obtained using feature-based approaches and traditional classifiers, with accuracy scores above 70%, and analyze the shortcomings of our current results and avenues for further improvement, taking into account the intended use of our classifiers in helping human officers to cope with thousands of yearly complaints.

## 1 Introduction

Artificial intelligence (AI) techniques are nowadays widespread in virtually every sector of human activity. Not only the private sector but also public administration institutions and governments are looking into ways of taking advantage of AI to deal with their specific problems and exploit their substantial amounts of both structured and unstructured information. Natural language processing (NLP) techniques are being employed in this regard to handle text available in the web (such as in social networks or newswires) and, most importantly, written forms of direct interaction between citizens and governmental institutions (Eggers, 2019).

Several governmental institutions provide public services electronically. Moreover, such institutions are responsible for processing citizen feedback (such as requests or complaints), often materialized through email or contact forms in so-called virtual counters. The amount of such contacts can become intractable in a short period of time, depending on the size of the country/administrative region. Based on such information, NLP techniques can be used to improve public services (Kowalski et al., 2019).

This paper focuses on the needs of the Portuguese Economic and Food Safety Authority (ASAE)[1], a national administrative authority specialized in the context of food safety and economic surveillance, responsible for monitoring and enforcing regulatory legislation. One of the main inputs of this institution is comprised of citizen complaints on the activity of economic agents, with more than twenty thousand complaints being received annually. Usually, more than 30% of these are found not to be in the jurisdiction of this authority; the remaining are sent to specific operational units. The use of human labor to analyze and properly handle these complaints is a bottleneck, bringing the need to automate this process to the extent possible. Doing it effectively is hindered by the fact that contact forms typically include free-form text fields, bringing high variability to the quality of the content written by citizens (which can be considered as user-generated content (Momeni et al., 2015)).

In this paper we present an analysis of a rich dataset containing 150,700 complaints related to food safety and economic surveillance. We also present machine learning-based classifiers that perform accurately for three key dimensions that are especially important for ASAE. Initial experiments using Deep Learning architectures are also reported. To the best of our knowledge, this is

[1] http://www.asae.gov.pt/welcome-to-website-asae.aspx

*Proceedings of the Second Workshop on Economics and Natural Language Processing*, pages 51–60
Hong Kong, November 4. ©2019 Association for Computational Linguistics

the first study of its kind regarding food safety and economic surveillance complaints for the Portuguese language.

In Section 2, we start by providing a short analysis of related work. Section 3 explains the overall complaint processing steps considered and provides an exploratory data analysis. Section 4 explains the main choices regarding preprocessing and feature extraction, that are common to all addressed classification tasks, whose details and experimental results are further developed in Sections 5, 6 and 7. Section 8 concludes the paper and points to directions for future work.

## 2 Related Work

Works on analyzing user-generated content mostly study social media data (Batrinca and Treleaven, 2015), focusing on tasks such as sentiment analysis (Eshleman and Yang, 2014; Forte and Brazdil, 2016) and opinion mining (Petz et al., 2013), or predicting the usefulness of product reviews (Diaz and Ng, 2018). For instance, Forte and Brazdil (2016) focus on sentiment polarity of Portuguese comments from the customer service department of a major Portuguese telecommunications company and use a lexicon-based approach enriched with domain-specific terms, formulating specific rules for negation and amplifiers.

Literature on non-social media complaint analysis is considerably more scarce, mainly due to the fact that such data is typically not publicly available. Even so, the problem has received significant attention from the NLP community, as a recent task on consumer feedback analysis shows (Liu et al., 2017). Given the different kinds of analysis one may want to undertake, however, the task concentrates on a single goal: to distinguish between comment, request, bug, complaint, and meaningless. In our work, we need to further analyze the contents of complaints, with a finer granularity.

Ordenes et al. (2014) propose a framework for analyzing customer experience feedback, using a linguistics-based model. This approach explores the identification of activities, resources and context, so as to automatically distinguish compliments from complaints, regarding different aspects of customer feedback. The work focuses on a single activity domain and, in the end, aims at obtaining a refined sentiment analysis model. In our work, we avoid entering into a labor-intensive annotation process of domain-specific data and focus on cross-domain classification tasks that help in complaint processing.

Traditional approaches to text categorization employ feature-based sparse models, using bag-of-words and Term Frequency-Inverse Document Frequency (TF-IDF) encoding. In the context of insurance complaint handling, Dong and Wang (2015) make use of synonyms and Chi-square statistics to reduce the dimensionality of the feature space. More recent techniques, such as word embeddings (Mikolov et al., 2013) and recurrent neural networks (RNNs) (Elman, 1990), have also been used in complaint classification. Assawinjaipetch et al. (2016) employ these methods to classify complaints of a single company into one of nine classes, related to the specific aspect that is being criticized.

Given the noisy nature of user-generated content, dealing with complaints as a multi-label classification problem can be effective, even when the original problem is single-labeled. Ranking algorithms (Li, 2014; Momeni et al., 2015) are a promising approach in this regard, providing a set of predictions sorted by confidence. These techniques have been applied in complaint analysis by Fauzan and Khodra (2014), although with modest results.

Kalyoncu et al. (2018) approach customer complaint analysis from a topic modeling perspective, using techniques such as Latent Dirichlet Allocation (LDA) (Blei et al., 2003). This work is not so much focused on automatically processing complaints, but instead on providing a visualization tool for mobile network operators.

## 3 Complaint Data

Among several other responsibilities pertaining to economic and food safety, ASAE, the Portuguese Economic and Food Safety Authority, is also responsible for handling consumer complaints. These complaints can be submitted by any citizen, either through a website form submission (including a free-form text field) or directly by email. Once a complaint is received, it must be handled by an officer, who is responsible for extracting all relevant information and filling it as part of a more structured complaint format in the back-end. This structured complaint will then be used to decide if and when it should be investigated.

### 3.1 Key Dimensions

There are a number of fields that are part of the final complaint structure before it is acted upon. More specifically, and in addition to context information such as names and addresses of the entities involved, there are three key dimensions.

The first is the type of *economic activity* related to the complaint. In total 11 categories can be assigned to a complaint ranging, for example, from online sales to restaurants. The type of activity is an important aspect for ASAE coordination, as a number of its operations are dedicated to specific activities within a long-term predefined strategic plan.

The second key dimension is *infraction severity*. This dimension concerns the infractions implied by the complaint. Each infraction can be considered an administrative infringement, a crime or a simple consumer conflict. Understanding the severity of infractions allows ASAE to prioritize investigating more serious and potentially harmful complaint targets.

Finally, the third key dimension is *competence*. This dimension essentially indicates whether a complaint refers to an event that is within ASAE jurisdiction, or if it should be treated by a different judicial or governmental entity. This distinction is important because ASAE should not investigate complaints outside its jurisdiction and should also forward the complaint to the competent authority.

### 3.2 Exploratory Data Analysis

The dataset used for the experiments presented in this work consists of 150,700 complaints, written in Portuguese, received by ASAE over the course of 11 years, starting in 2008 and ending in 2018. In addition to the textual contents of each complaint, the dataset contains all annotations performed by ASAE officers. This allows for a detailed analysis of the complaints received by the public entity, which falls outside the scope of this paper but is summarized in this section.

Table 1 shows the distribution for economic activities. It is fairly unbalanced, with a majority class taking 32.07% of all examples, and the most underrepresented class having only 0.02%. The top 3 classes represent in total 72% of the dataset. Class Z is a special case because it signals that no economic activity has been perceived in the complaint. Only 146,847 complaints are considered for this dimension because the remaining 3,853 do not have a valid economic activity label, i. e., differently from class Z examples which indicate that no economic activity was identified, these examples do not have a classification label in terms of economic activity.

Each complaint can include several different infraction indications, which in turn means one complaint can contain infractions of varying severity. In order to simplify the problem, we decided to focus on the highest infraction severity implied by each complaint. This makes prioritization easier – a complaint indicating crime is more severe than a complaint pointing only to administrative infringements – but also makes classification fuzzier due to the overlap between crimes and administrative infringements in some cases. The distribution among the resulting three classes is shown in Table 2.

Table 3 shows the data distribution based on the competence label. While the original dataset provides a list of entities that should ultimately handle each complaint, the focus of the experiments reported in this paper is solely to determine whether ASAE is one of them.

| Class | # compl | % |
|---|---:|---:|
| I - Primary Production | 572 | 0.39 |
| II - Industry | 4,214 | 2.87 |
| III - Restoration | 47,098 | 32.07 |
| IV - Wholesalers | 631 | 0.43 |
| V - Retail | 13,904 | 9.47 |
| VI - Direct selling | 27 | 0.02 |
| VII - Distance selling | 4,760 | 3.24 |
| VIII - Production & Trade | 14,236 | 9.69 |
| IX - Service Providers | 35,737 | 24.34 |
| X - Safety & Environment | 1,905 | 1.30 |
| Z - No activity identified | 23,763 | 16.18 |
| Total | 146,847 | 100.00 |

Table 1: Economic activity class distribution

| Class | # compl | % |
|---|---:|---:|
| Crime | 8,086 | 5.37 |
| Admin. infringement | 69,012 | 45.79 |
| Other | 73,602 | 48.84 |
| Total | 150,700 | 100.00 |

Table 2: Infraction severity class distribution

The complaints are evenly distributed across time, roughly 14,000 per year, with a slight increase towards the last 5 years. A geographical

| Class | # compl | % |
|---|---:|---:|
| ASAE and others | 94,140 | 62.47 |
| Other | 56,560 | 37.53 |
| Total | 150,700 | 100.00 |

Table 3: Competence class distribution (binary setting)

analysis reveals that more densely populated areas generate more complaints, as expected.

A majority of 63% complaints are received via the ASAE website. The complaint form is mostly free-text but it does specifically request the author to identify himself by providing his name, address, phone number and email address. The author is also requested to identify the entity targeted by the complaint using the same information. Unfortunately, not every complaint provides enough context or information to successfully determine the target entity, making it impossible to investigate.

## 4 Experimental Setup

In order to implement machine learning classifiers based on the textual contents of each complaint, and given their user-generated content nature, a previous preprocessing step was necessary. Based on an earlier work that tackled economic activity prediction on a smaller sample of this dataset (Barbosa et al., 2019), the dataset was preprocessed using the Natural Language Toolkit (NLTK) (Bird et al., 2009) to perform tokenization, lemmatization and remove stop words from Portuguese text. Furthermore, from among the different feature-based representations explored by Barbosa et al. (2019), a TF-IDF weighted vector was found to be the most effective method of representing each document. TF-IDF outperformed fastText-based (Joulin et al., 2016) and BERT-based (Devlin et al., 2018) representations, using traditional machine learning approaches, specifically Support Vector Machines (SVM) (Cortes and Vapnik, 1995).

For all experiments reported in this paper, the split between training and test sets was performed bearing in mind that the processes used by ASAE have suffered small changes over the last decade and that the ultimate goal is to help officers perform their work more efficiently when handling complaints nowadays. As such, the test set used in these experiments has been drawn from the last 5 years of data only. This also ensures the results for the task of economic activity prediction reported in this paper can be compared to results from earlier work (Barbosa et al., 2019), which were obtained using the same test set. A total of roughly 25,000 examples make up this test set, 16% of all available data. For each task, a different stratified splitting was performed, to ensure that the resulting test sets followed the target distribution.

The following classifiers were employed: Naïve Bayes (NB) (Manning et al., 2008), K-Neighbors (Altman, 1992), SVM, Stochastic Gradient Descent (SGD) (Zhang, 2004), Decision Tree Classifier (Quinlan, 1986), Randomized Decision Trees (also know as extra-trees)(Geurts et al., 2006), Random Forests (Breiman, 2001), and Bagging Classifier (Breiman, 1996). For all ensemble models (i.e. Randomized Decision Trees, Random Forests, and Bagging Classifier), Decision Trees are used as weak classifiers with default parameters. For reference, we also report the scores of a random classifier that generates predictions based only on the training set label distribution (dubbed “Random (stratified)”).

The scikit-learn library (Pedregosa et al., 2011) was used to implement bag-of-words and TF-IDF encoding, train-test set stratified splitting and all classifiers, unless otherwise stated.

As evaluation metric, we focused on the accuracy score (Acc), because, for the application scenario at ASAE, we aim to classify the complaints as accurately as possible. However, given the unbalanced nature of the label distribution, we also report Macro-F1 scores, which provide an estimate on how good the classifiers are across different labels, without taking into account label imbalance.

## 5 Economic Activity Prediction

One of the first steps needed to analyze a complaint concerns the identification of the targeted economic activity, from those shown in Table 1. We model this as a classification problem with 11 classes. Given the relatively high number of classes, we also look at the performance of each classifier considering its ranked output. This approach is aligned with the potential usage of the classifier, which is meant to help humans analyze complaints by providing likely classification labels (as opposed to imposing a definitive one).

Table 4 summarizes the scores obtained for this task, where Acc@$k$ and Macro-F1@$k$ are accuracy and macro-F1 scores, respectively, when considering that the classifier has made a correct prediction if any of the $k$ most confidently predicted classes (top-$k$) corresponds to the target label. Overall, the best classifier is a SVM with a linear kernel, achieving the highest accuracy and macro-F1 scores for every top-$k$, with the exception of top-3 accuracy, where SGD outperforms SVM by under 1%. Both SVM and SGD perform considerably better than any other alternatives, notably Random Forests. All classifiers significantly outperform the stratified random baseline.

### 5.1 Error Analysis

Based on the different accuracy and average macro-F1 scores obtained, we have decided to focus on SVM for the sake of error analysis. The SVM confusion matrix is shown in Table 5 and is complemented by the per-class precision and recall metrics displayed in Table 6.

The influence of majority classes III and IX is visible, while class Z (in which no economic activity is identified) seems to be the most ambiguous for the classifier, given also its high number of examples. In fact, class III has the highest recall, but also precision. Most other classes have good precision scores, while some of them suffer from low recall, namely: classes I, IV, and X. Class VI contains too few examples to be considered.

While inspecting some of the misclassified instances, a number of issues became apparent. Some examples comprise short text complaints, not providing enough information to classify their target economic activity. A small number of complaints are not written in Portuguese. Some complaint texts are followed by non-complaint-related content, sometimes in English. Some classes exhibit semantic overlap. For instance, class VIII (Production & Trade) overlaps with classes II (Industry) and V (Retail). That means that complaints labeled VII often contain words that are highly correlated with II and V. A non-negligible number of examples refer to previously submitted complaints, either to provide more data or to request information on their status. These cases do not contain the complaint itself, the same happening when a short text simply includes meta-data or points to an attached file. Finally, we were able to identify some complaints that have been misclassified by the human operator.

As mentioned previously, and plainly observable in Table 1, this classification problem is very imbalanced. In previous work (Barbosa et al., 2019), while considering a sample of the dataset with half the time window (and thus with approximately half the size, while maintaining a similar class distribution), we have tried employing both random undersampling and random oversampling (He and Garcia, 2009), in order to improve the overall classification performance and, more specifically, the performance on minority classes. However, such attempts did not succeed, consistently worsening results.

Because class Z is used to indicate that no activity has been identified and, for that reason, is highly diffuse, we have conducted a few experiments to try to find better approaches of dealing with this class. Removing class Z from the training subset, while assuming this class as the correct label in the absence of an above-threshold confidence in any class, did not bring satisfactory results, as no appropriate threshold could be found. Otherwise, assuming class Z as the correct label when it is one of the top-2 predicted classes also lowered scores significantly.

### 5.2 Deep Learning Approaches

As part of an effort to further improve the classification results on this task, that proved to be more challenging given the number of classes and their similarities, a shift was made from traditional feature-based approaches to word embeddings and deep neural network architectures (deep learning approaches). In particular, a number of experiments using long short-term memory neural networks (LSTM) (Hochreiter and Schmidhuber, 1997) were performed. While these results are preliminary, the best configuration of an LSTM-based classifier achieved an accuracy of 0.695 and a macro-F1 of 0.44. This particular configuration used a hidden layer of size 1024 and we retrain the embeddings with 300 dimensions that were initialized randomly. Adam (Kingma and Ba, 2014) was used for optimization and negative log-likelihood loss chosen as the cost function. Standard first choices were used for the remaining hyperparameters, including: learning rate of 0.001 (Kingma and Ba, 2014), dropout of 0.2 (Srivastava et al., 2014), and batch size of 32. Initial experiments focused on variations of these parameters: learning rates between 0.001 and 0.0001; dropout between 0.2 and 0.5. Runs with fixed or trainable embeddings and different hidden layer sizes (128 to 1024) were also attempted.

| Classifier | Acc@1 | Acc@2 | Acc@3 | Macro-F1@1 | Macro-F1@2 | Macro-F1@3 |
|---|---|---|---|---|---|---|
| Random (stratified) | 0.2035 | 0.3314 | 0.3704 | 0.09 | 0.16 | 0.23 |
| Bernoulli NB | 0.4554 | 0.6567 | 0.7998 | 0.16 | 0.28 | 0.37 |
| Multinomial NB | 0.4786 | 0.6049 | 0.7332 | 0.11 | 0.18 | 0.28 |
| Complement NB | 0.5922 | 0.7873 | 0.8944 | 0.29 | 0.48 | 0.61 |
| K-Neighbors | 0.2949 | 0.4701 | 0.6078 | 0.18 | 0.31 | 0.42 |
| SVM (linear) | **0.7554** | **0.8792** | 0.9320 | **0.57** | **0.72** | **0.79** |
| SGD | 0.7379 | 0.8739 | **0.9404** | 0.51 | 0.66 | 0.75 |
| Decision Tree | 0.5987 | 0.6985 | 0.7141 | 0.39 | 0.46 | 0.48 |
| Extra Tree | 0.4162 | 0.5265 | 0.5495 | 0.26 | 0.32 | 0.35 |
| Random Forests | 0.6247 | 0.7854 | 0.8807 | 0.37 | 0.50 | 0.59 |
| Bagging | 0.6617 | 0.8054 | 0.8709 | 0.44 | 0.56 | 0.64 |

Table 4: Economic activity prediction results

| | | Predicted | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | I | II | III | IV | V | VI | VII | VIII | IX | X | Z |
| Actual | I | **43** | 5 | 4 | 0 | 11 | 0 | 0 | 4 | 4 | 0 | 17 |
| | II | 0 | **384** | 127 | 6 | 61 | 0 | 1 | 15 | 30 | 0 | 78 |
| | III | 0 | 68 | **7,036** | 2 | 89 | 0 | 5 | 40 | 209 | 4 | 205 |
| | IV | 0 | 10 | 10 | **26** | 17 | 0 | 0 | 6 | 8 | 0 | 30 |
| | V | 1 | 22 | 141 | 5 | **1,845** | 0 | 9 | 56 | 63 | 0 | 113 |
| | VI | 0 | 0 | 0 | 0 | 0 | **0** | 0 | 1 | 0 | 0 | 0 |
| | VII | 0 | 0 | 7 | 0 | 12 | 0 | **864** | 58 | 114 | 0 | 101 |
| | VIII | 1 | 7 | 122 | 4 | 57 | 0 | 49 | **1,502** | 268 | 6 | 259 |
| | IX | 1 | 9 | 379 | 4 | 35 | 0 | 53 | 166 | **4,895** | 6 | 380 |
| | X | 0 | 4 | 22 | 1 | 11 | 0 | 13 | 62 | 83 | **114** | 56 |
| | Z | 11 | 65 | 480 | 9 | 155 | 0 | 108 | 356 | 824 | 25 | **1,388** |

Table 5: Economic activity prediction confusion matrix using SVM (top-1)

The training process was allowed to run for a maximum of 20 epochs. However, for each epoch, the training process measured accuracy on a separate development set and kept the model that performed best. The neural network architectures were implemented using PyTorch (Paszke et al., 2017).

While the results are still far from the accuracy obtained using SVMs, 0.755, further experiments are planned using pre-trained embeddings, such as fastText and BERT, combined with different deep learning architectures, including convolutional neural networks (Dos Santos and Gatti, 2014) and attention mechanisms (Bahdanau et al., 2015; Yang et al., 2016).

## 6 Infraction Severity Prediction

The priority of a complaint is directly related to the infractions that emerge from the reported information. Instead of predicting infractions, however, we focus on their severity, in a three-layered framework (as shown in Table 2). As mentioned in Section 3, we decided to reduce the problem from a multi-label and multi-class setting to a single-label problem, where we identify the most severe type of infraction evidenced by the complaint: a crime or an administrative infringement.

The accuracy and macro-F1 scores obtained using different classifiers are shown in Table 7. Contrary to the results of predicting economic activity, SGD performs slightly better in terms of accuracy, while SVM still leads on macro-F1 score. Once again, both SVM and SGD outperform other classifiers. However, for this task the differences are not as pronounced, especially in relation to Bagging and to a lesser extent Random Forests. Every classifier outperforms the baseline.

### 6.1 Error Analysis

As before, we focus on SVM for the sake of error analysis, although SGD would also be a valid op-

| | Precision | Recall |
|---|---|---|
| I | 0.75 | 0.49 |
| II | 0.67 | 0.55 |
| III | 0.84 | 0.92 |
| IV | 0.46 | 0.24 |
| V | 0.80 | 0.82 |
| VI | – | 0.00 |
| VII | 0.78 | 0.75 |
| VIII | 0.66 | 0.66 |
| IX | 0.75 | 0.83 |
| X | 0.74 | 0.31 |
| Z | 0.53 | 0.41 |

Table 6: Economic activity prediction precision and recall per class (top-1)

| Classifier | Acc | Macro-F1 |
|---|---|---|
| Random (stratified) | 0.4499 | 0.33 |
| Bernoulli NB | 0.5909 | 0.40 |
| Multinomial NB | 0.6731 | 0.46 |
| Complement NB | 0.6750 | 0.50 |
| K-Neighbors | 0.4859 | 0.36 |
| SVM (linear) | 0.7075 | **0.66** |
| SGD | **0.7231** | 0.64 |
| Decision Tree | 0.6242 | 0.56 |
| Extra Tree | 0.5709 | 0.47 |
| Random Forests | 0.6881 | 0.55 |
| Bagging | 0.6805 | 0.62 |

Table 7: Infraction severity prediction results

tion. By analyzing the confusion matrix shown in Table 8, it is possible to observe that class "Administrative infringement" and "Others" have a considerable number of cases where the prediction is swapped. Furthermore, several crime cases are being wrongly classified. A source of confusion between administrative infringements and crimes is their co-occurrence in some complaints of the original data (as mentioned in Section 3.2), and results from reducing the problem to a single-label setting.

| | | Predicted | | |
|---|---|---|---|---|
| | | Crime | Adm. infr. | Other |
| Actual | Crime | **579** | 362 | 324 |
| | Adm. infr. | 95 | **8,371** | 3,089 |
| | Other | 153 | 2,984 | **8,000** |

Table 8: Infraction severity prediction confusion matrix using SVM

Although the accuracy and macro-F1 scores are not low, there is considerable room for improvement in this particular task. Taking into account the application of this classification model in food safety and economic surveillance, special attention should be given to false negatives of the "Crime" and "Administrative infringement" classes.

## 7 Competence Prediction

In practice, identifying the competent entity(ies) to handle a complaint is determined by the output of the previous two dimensions: economic activity and infractions. However, since we are not directly predicting infractions (but rather their severity), we have chosen to predict the competence directly from the complaint contents. As mentioned in Section 3.2, we decided to reduce the competence prediction problem to a binary classification setting (as per Table 3), where we identify whether ASAE is one of the institutions responsible to handle the complaint or not.

The accuracy and macro-F1 scores obtained using different classifiers are shown in Table 9. In consistence with the other tasks, SGD and SVM perform better than all remaining classifiers, with Bagging and Random Forests slightly behind. For this task, K-Neighbours and Multinomial NB are not particularly far from the baseline.

| Classifier | Acc | Macro-F1 |
|---|---|---|
| Random (stratified) | 0.5308 | 0.50 |
| Bernoulli NB | 0.6866 | 0.65 |
| Multinomial NB | 0.6661 | 0.53 |
| Complement NB | 0.6929 | 0.60 |
| K-Neighbors | 0.5877 | 0.57 |
| SVM (linear) | **0.7953** | **0.78** |
| SGD | 0.7927 | **0.78** |
| Decision Tree | 0.7002 | 0.68 |
| Extra Tree | 0.6532 | 0.63 |
| Random Forests | 0.7477 | 0.70 |
| Bagging | 0.7440 | 0.73 |

Table 9: Competence prediction results

### 7.1 Error Analysis

SVM is again chosen for error analysis. Table 10 presents the confusion matrix for this task and shows there is a considerable amount of cases where the prediction is incorrect. As with the previous task, we are particularly interested in addressing false negatives of the ASAE class.

| | | Predicted | |
|---|---|---|---|
| | | ASAE | Other |
| Actual | ASAE | **12,408** | 2,243 |
| | Other | 2,662 | **6,644** |

Table 10: Competence prediction confusion matrix using SVM

It should be noted that our results show that it is possible, to a large extent, to derive ASAE's competence directly from the complaint text (with a recall of 85%). Albeit this does not correspond to the current practice, it does comprise a promising shortcut to this task.

## 8 Conclusions

In this paper, we present our findings regarding the classification of complaints, written in the Portuguese language, along three key dimensions: economic activity, infraction severity and competence. Traditional machine learning and natural language processing approaches, such as bag-of-words with TF-IDF encoding and SVM models, provide fairly accurate classifiers for these tasks. Our preliminary work using Deep Learning approaches requires further investigation (*e.g.* exploring different architectures) and have yet to reach the same levels of performance.

This work can be integrated in an AI-powered web platform to help ASAE officers in their efforts to tackle the large amount of complaints received, not only by providing semi-automatic annotating capabilities but also for managing work prioritization. The classifiers, however, still reveal some limitations. In particular, for economic activity, the Z class – no discernible economic activity – is still a source of considerable confusion. Strategies to overcome this limitation have not been successful yet. For infraction severity, it would be important to achieve better results distinguishing crimes from other infractions, as these should receive the highest priority.

Additional work is planned to counter these limitations and strive for more accurate classifiers, in an effort to further improve the performance of the system. In particular, we are experimenting with different deep learning architectures, pre-trained word embeddings, and hyperparameter fine-tuning of the machine learning models.

## Acknowledgments

This work is supported by project IA.SAE, funded by Fundação para a Ciência e a Tecnologia (FCT) through program INCoDe.2030. Gil Rocha is supported by a PhD scholarship from FCT (SFRH/BD/140125/2018).

## References

N. S. Altman. 1992. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):175–185.

Panuwat Assawinjaipetch, Kiyoaki Shirai, Virach Sornlertlamvanich, and Sanparith Marukata. 2016. Recurrent Neural Network with Word Embedding for Complaint Classification. In *Proceedings of the Third International Workshop on Worldwide Language Service Infrastructure and Second Workshop on Open Infrastructures and Analysis Frameworks for Human Language Technologies (WLSI/OIAF4HLT2016)*, pages 36–43, Osaka, Japan. The COLING 2016 Organizing Committee.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Luís Barbosa, João Filgueiras, Gil Rocha, Henrique Lopes Cardoso, Luís Paulo Reis, João Pedro Machado, Ana Cristina Caldeira, and Ana Maria Oliveira. 2019. Automatic Identification of Economic Activities in Complaints. In *Statistical Language and Speech Processing, 7th International Conference, SLSP 2019, Ljubljana, Slovenia, October 1416, 2019, Proceedings*, volume 11816 of *LNAI*. Springer.

Bogdan Batrinca and Philip C. Treleaven. 2015. Social media analytics: a survey of techniques, tools and platforms. *AI & Society*, 30(1):89–116.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*, 1st edition. O'Reilly Media, Inc.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022.

Leo Breiman. 1996. Bagging predictors. *Machine Learning*, 24(2):123–140.

Leo Breiman. 2001. Random forests. *Machine Learning*, 45(1):5–32.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning*, 20(3):273–297.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Gerardo Ocampo Diaz and Vincent Ng. 2018. Modeling and Prediction of Online Product Review Helpfulness: A Survey. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 698–708.

Shuang Dong and Zhihong Wang. 2015. Evaluating Service Quality in Insurance Customer Complaint Handling throught Text Categorization. In *2015 Int. Conf. on Logistics, Informatics and Service Sciences (LISS)*, pages 1–5. IEEE.

Cicero Dos Santos and Maira Gatti. 2014. Deep convolutional neural networks for sentiment analysis of short texts. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 69–78.

William D. Eggers. 2019. Using AI to unleash the power of unstructured government data. *Deloitte Insights*.

Jeffrey L. Elman. 1990. Finding structure in time. *Cognitive Science*, 14(2):179–211.

Ryan M. Eshleman and Hui Yang. 2014. "Hey #311, Come Clean My Street!": A Spatio-temporal Sentiment Analysis of Twitter Data and 311 Civil Complaints. In *2014 IEEE Fourth International Conference on Big Data and Cloud Computing*, pages 477–484.

Ahmad Fauzan and Masayu Leylia Khodra. 2014. Automatic Multilabel Categorization using Learning to Rank Framework for Complaint Text on Bandung Government. In *2014 Int. Conf. of Advanced Informatics: Concept, Theory and Application (ICAICTA)*, pages 28–33. Institut Teknologi Bandung, IEEE.

Ana Catarina Forte and Pavel B. Brazdil. 2016. Determining the Level of Clients' Dissatisfaction from Their Commentaries. In *Computational Processing of the Portuguese Language - 12th Int. Conf., PROPOR 2016*, volume 9727 of *Lecture Notes in Computer Science*, pages 74–85. Springer.

Pierre Geurts, Damien Ernst, and Louis Wehenkel. 2006. Extremely randomized trees. *Machine Learning*, 63(1):3–42.

Haibo He and Edwardo A. Garcia. 2009. Learning from Imbalanced Data. *IEEE Trans. on Knowl. and Data Eng.*, 21(9):1263–1284.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Feyzullah Kalyoncu, Engin Zeydan, Ibrahim Onuralp Yigit, and Ahmet Yildirim. 2018. A Customer Complaint Analysis Tool for Mobile Network Operators. In *2018 IEEE/ACM Int. Conf. on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 609–612. IEEE.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Radoslaw Kowalski, Marc Esteve, and Slava Jankin Mikhaylov. 2019. Improving Public Services by Mining Citizen Feedback: An Application of Natural Language Processing. *EasyChair preprint 1103*.

Hang Li. 2014. *Learning to Rank for Information Retrieval and Natural Language Processing*, 2 edition. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publ., San Rafael, CA.

Chao-Hong Liu, Yasufumi Moriya, Alberto Poncelas, and Declan Groves. 2017. IJCNLP-2017 Task 4: Customer Feedback Analysis. In *Proceedings of the IJCNLP 2017, Shared Tasks*, pages 26–33, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

Elaheh Momeni, Claire Cardie, and Nicholas Diakopoulos. 2015. A Survey on Assessment and Ranking Methodologies for User-Generated Content on the Web. *ACM Computing Surveys*, 48(3):41:1–41:49.

Francisco Villarroel Ordenes, Babis Theodoulidis, Jamie Burton, Thorsten Gruber, and Mohamed Zaki. 2014. Analyzing Customer Experience Feedback Using Text Mining: A Linguistics-Based Approach. *Journal of Service Research*, 17(3):278–295.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch. In *NIPS Autodiff Workshop: The Future of Gradient-based Machine Learning Software and Techniques*.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Gerald Petz, Michał Karpowicz, Harald Fürschuß, Andreas Auinger, Václav Stříteský, and Andreas Holzinger. 2013. Opinion Mining on the Web 2.0 – Characteristics of User Generated Content and Their Impacts. In *Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data*, pages 35–46. Springer.

J. R. Quinlan. 1986. Induction of decision trees. *Mach. Learn.*, 1(1):81–106.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1480–1489.

Tong Zhang. 2004. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *Proceedings of the Twenty-first International Conference on Machine Learning*, ICML '04, pages 116–, New York, NY, USA. ACM.