Text Representation through Multimodal Variational Autoencoder for One-Class Learning

Marcos Paulo Silva Gôlo[1], **Ricardo Marcondes Marcacini**[1]

[1]Institute of Mathematics and Computer Sciences, USP, São Carlos, SP, Brazil

Abstract. *Multi-class learning (MCL) methods perform Automatic Text Classification (ATC), which requires labeling for all classes. MCL fails when there is no well-defined information about the classes and requires a great effort to label instances. One-Class Learning (OCL) can mitigate these limitations since the training only has instances from one class, reducing the labeling effort and making the ATC more appropriate for open-domain applications. However, OCL is more challenging due to the lack of counterexamples for model training, requiring more robust representations. However, most studies use unimodal representations, even though different domains contain other information that can be used as modalities. Thus, this study proposes the Multimodal Variational Autoencoder (MVAE) for OCL. MVAE is a multimodal method that learns a new representation from more than one modality, capturing the characteristics of the interest class in an adequate way. MVAE explores semantic, density, linguistic, and spatial information modalities. The main contributions are: (i) a multimodal method for ATC through OCL; (ii) MVAE for fake news detection; (iii) relevant reviews detection via MVAE; and (iv) sensing events through MVAE.*

1. Introduction

ATC assigns a previously defined label in unlabeled textual documents. MCL is a strategy for ATC. In MCL, the user must know all classes of the problem and label documents for all those classes in the training step. Therefore, the user must label documents of classes even if he/she is not interested, implying two limitations: when the user does not label examples for all domain classes and when a new domain class comes up. Thus, MCL may not be viable when there is no well-defined knowledge of all classes and requires a greater effort to label the instances for each class [Aggarwal 2018].

One approach that mitigates some limitations presented by the MCL is One-Class Learning (OCL) [Alam et al. 2020, Tax 2001]. OCL uses only examples from one class (interest class) to learn, i.e., the learning is in the absence of counterexamples [Tax 2001]. OCL will be able to identify whether an instance belongs to the interest class, reducing the labeling effort and being more appropriate for open-domain applications or applications in which the user is interested in one class [Gôlo et al. 2021a, Gôlo et al. 2021b].

The learning process is more challenging for the OCL due to the lack of counterexamples, requiring more robust representations for text. Generally, studies use the traditional Bag-of-Words (BoW) technique [Manevitz and Yousef 2007, Junior and Rossi 2017]. Other studies explore dimensionality reduction techniques [Kumar and Ravi 2017b, Gôlo et al. 2019]. Finally, studies used language models via neural networks [Ruff et al. 2019, Mayaluru 2020]. This study highlights that all these models are unimodal. However, the text domains have different useful information to be used as distinct data modalities [Zhou et al. 2020, Guo et al. 2019].

Despite the benefits of representations generated through multimodal learning [Li et al. 2018, Guo et al. 2019], using multimodality to represent texts in the OCL scenario is a gap in the literature. Thus, this study has two research challenges related to multimodal learning. First, **Multimodal text representations for OCL**: a representation suitable for OCL in which interest texts are closer to each other while far from non-interest texts. Although fusing multiple modalities allows for more robust representations, most ATC through OCL studies explore only unimodal text representation methods [Gôlo et al. 2019, Mayaluru 2020]. Second, **Unsatisfactory OCL performances with few labeled instances**: even though OCL decreases the user's labeling effort, the fewer labeled instances, the lesser the user's effort. The reduction of training instances can harm OCL performance. A research challenge involves investigating appropriate representations to reduce dependence on large interest training sets, preserving OCL performance.

Given the research challenges, this study proposed the Multimodal Variational Autoencoder (MVAE) for OCL. MVAE is a representation learning method that learns a representation from multiple modalities through a neural network based on VAEs that are generative models considered one of the state-of-the-art for text representation learning [Xu and Durrett 2018, Wang et al. 2019]. Moreover, another research goal is to analyze the proposal in real-world applications with few labeled instances, such as detecting fake news, relevant reviews, and events, to verify the generalizability in different domains.

This study has two main contributions. First, **MVAE for OCL in ATC**: the study proposes an MVAE architecture that generates more suitable textual representations for OCL. In particular, our MVAE explores as modalities: (i) pre-trained embeddings from the BERT multilingual to incorporate more semantic knowledge; (ii) topic information from the high-density regions; (iii) features with the linguistic structure of the texts; and (iv) geolocation data (latitude and longitude). Our MVAE also proved robust for scenarios with few labeled instances in three domains, further reducing the labeling effort.

Second, **The study applies MVAE in three real scenarios**. First, the study detects fake news through OCL with the MVAE representations [Gôlo et al. 2021a]. The MVAE learns a new representation from the combination of promising modalities: text embeddings, linguistic features, and topic/density information. Second, the study detects relevant app reviews through OCL with the MVAE representations [Gôlo et al. 2022]. Our MVAE explores text embeddings and high-density regions through review topics or subtopics. Third, the study detects events of interest through OCL with the MVAE representations [Gôlo et al. 2021b]. MVAE represents the events learning a unified representation from text embeddings, geolocation, and density modalities. The results in all scenarios show that the MVAE with 3% of labeled instances outperforms other representation methods with more labeling.

2. Related Work

Bag-of-Words. Pioneering studies for one-class text learning used the bag-of-words (BoW) to represent the texts [Manevitz and Yousef 2001, Manevitz and Yousef 2007]. The studies have some limitations, such as the lack of k-fold cross-validation, the use of only one dataset, and the lack of statistical tests. In addition, other studies after the pioneers used the BoW with different term weights to improve the representation [Junior and Rossi 2017, Gôlo et al. 2019] without the limitations presented above. Fi-

nally, the BoW has limitations, such as high dimensionality and sparsity and inefficiency in the presence of synonyms and ambiguity. Thus, one-class text learning evolved to the use of dimensionality reduction techniques.

Dimensionality Reduction. Given the limitations of the BoW, studies used dimension reduction techniques, generating non-sparse and lower dimension representations to improve the representation and, consequently, the ATC [Kumar and Ravi 2017b, Kumar and Ravi 2017a, Gôlo et al. 2019]. The studies [Kumar and Ravi 2017b] and [Kumar and Ravi 2017a] have limitations, such as little/no variation of ATC parameters, use of few collections, lack of k-fold cross-validation, and use only recall/precision for evaluation. Furthermore, [Gôlo et al. 2019] concludes that compared with the BoW, the reduction techniques did not improve the ATC. Thus, one-class text learning evolved to using language models based on neural networks.

Language Models Based on Neural Networks. From 2018 to 2021, studies have investigated language models based on neural networks, such as Word2Vec and Bidirectional Encoder Representations from Transformers (BERT), that generates more semantic representations through word embeddings [Ruff et al. 2019, Cichosz 2020, Mayaluru 2020]. These methods, mainly BERT, obtained state-of-the-art results for ATC through OCL. The study highlights that the related work generates a representation focused mainly on the text words or sentences. However, several domains can contain other information useful for learning. Different representations can be interpreted as distinct textual data modalities, such as topics, sentiment, temporal, geographic, and semantic information [Zhou et al. 2020, Guo et al. 2019]. Multimodal representation learning methods explore these different types of information to learn a more robust representation [Li et al. 2018, Guo et al. 2019]. In this sense, the study proposes the Multimodal Variational Autoencoder for One-Class Text Learning.

3. Multimodal Variational Autoencoder for One-Class Learning

The study proposes MVAE with different modalities for the texts. The main modalities were the DistilBERT [Devlin et al. 2019] representation and a proposed modality for density representation. Our proposed density modality is the density information extracted from the DistilBERT embeddings. The study uses this modality as a visual modality based on the different topics of the interest class. Specifically, our proposal explores clustering methods to identify high-density regions from texts in which the study interprets as topics. Then, the study extracts features from each cluster through statistical measures that describe the merits of the structure, such as cluster cohesion and separability.

Consider a clustering with k clusters, i.e., $\mathcal{D} = C_1 \cup C_2, \cup \cdots \cup C_k$, in which C_j is a cluster of documents, and $2 \leq k < m$. Then, the study applies the silhouette coefficient [Rousseeuw 1987] in order to measure if a document belongs to a single topic or contains mixed topics. The silhouette for a document $\boldsymbol{d}_i$ represented by the embeddings of BERT $\boldsymbol{\lambda}_i$ assigned to a cluster C_j is given by Equation 1. The silhouette values range from -1 to +1. A high value indicates that a document is well-matched to its cluster and weakly matched to neighboring clusters. The study represents the density information by concatenating silhouette coefficient values considering each document in different clustering settings. For instance, given u different clustering settings, i.e., the clustering settings have different values of k, the study performs the Density modality by Equation 2.

$$s(\boldsymbol{d}_i, k) = \frac{\beta(\boldsymbol{\lambda}_i) - \alpha(\boldsymbol{\lambda}_i)}{\max(\alpha(\boldsymbol{\lambda}_i), \beta(\boldsymbol{\lambda}_i))}, \quad (1) \qquad \begin{aligned} \boldsymbol{\delta}_i = \{&s(\boldsymbol{d}_i, k_1), s(\boldsymbol{d}_i, k_2), \ldots, \\ &\ldots, s(\boldsymbol{d}_i, k_u - 1), s(\boldsymbol{d}_i, k_u)\}, \end{aligned} \quad (2)$$

in which $\alpha(\boldsymbol{\lambda}_i)$ is the average distance of $\boldsymbol{\lambda}_i$ to all documents of cluster C_j, $\beta(\boldsymbol{\lambda}_i)$ is the average distance of a document $\boldsymbol{\lambda}_i$ to all documents of the closest cluster C_o, $o \neq j$, and $s(\boldsymbol{d}_i, k_j)$ is the silhouette of $\boldsymbol{d}_i$ in cluster setting with k_j clusters.

After obtaining the modalities in a structured way, the study must fuse them to combine them. In multimodal learning, early fusion is one of the most common types of fusion. The early fusion can be represented by combining modalities before the machine learning process. For instance, early fusion can be done using simple operators such as concatenation, addition, and multiplication [Katsaggelos et al. 2015]. An advantage of early fusion is using only one data representation in the learning process. On the other hand, early fusion has some challenges, such as dealing with different dimensions, scales, and levels of importance of each modality. However, to deal with these challenges, it is also possible to use more complex strategies than simple operators, such as operators based on neural networks [Gao et al. 2020].

For our proposed MVAE, the study chose the early fusion given its advantages and because the study solves its disadvantage through multimodal representation learned from neural networks. Therefore, the study uses dense layers with the same shape as the first layers of our MVAEs, and in the second layer, the study uses a merging layer that works as a fusion operator. Thus, the study can choose different operators for our early fusion and still use modalities with different sizes while our neural network learns the importance of each modality.

MVAE is a neural network with an encoder and decoder step, such as an Autoencoder. However, our MVAE is a Variational Autoencoder (VAE) variant. Thus, MVAE has useful properties for representation learning [Wang et al. 2019]. The MVAE performs a sampling step based on a previous distribution model to generate a bottleneck. Therefore, MVAE bias the learning through a prior informed distribution model, which is attractive for OCL because the representations generated by MVAE will preserve the main characteristics from the interest class representations and the model distribution, generating a region of the interest class [Xu and Durrett 2018, Wang et al. 2019].

If our modalities were $\boldsymbol{\lambda}_i$ and $\boldsymbol{\delta}_i$, MVAE assumes that $\boldsymbol{z}_i$ generates $\boldsymbol{\lambda}_i$ and $\boldsymbol{\delta}_i$ using Equation 3, in which $p(\boldsymbol{\lambda}_i, \boldsymbol{\delta}_i)$ is defined via Equation 4. Integrals are computationally intractable. Thus, MVAE uses variational inference, an approximation technique, to solve the limitation. Therefore, MVAE approximates $p(\boldsymbol{z}_i|\boldsymbol{\lambda}_i, \boldsymbol{\delta}_i)$ to $q(\boldsymbol{z}_i|\boldsymbol{\lambda}_i, \boldsymbol{\delta}_i)$ (treatable distribution) through the Kullback-Leibler (KL) divergence. Finally, MVAE optimizes the marginal likelihood ($p(\boldsymbol{\lambda}_i, \boldsymbol{\delta}_i)$) using the $\log$ of the marginal likelihood by Equation 5.

$$p(\boldsymbol{z}_i|\boldsymbol{\lambda}_i, \boldsymbol{\delta}_i) = \frac{p(\boldsymbol{\lambda}_i, \boldsymbol{\delta}_i|\boldsymbol{z}_i)p(\boldsymbol{z}_i)}{p(\boldsymbol{\lambda}_i, \boldsymbol{\delta}_i)}, \quad (3) \qquad p(\boldsymbol{\lambda}_i, \boldsymbol{\delta}_i) = \int p(\boldsymbol{\lambda}_i, \boldsymbol{\delta}_i|\boldsymbol{z}_i)p(\boldsymbol{z}_i)dz, \quad (4)$$

$$\log p_\Theta(\boldsymbol{\lambda}_i, \boldsymbol{\delta}_i) = KL(q_\Phi(\boldsymbol{z}_i|\boldsymbol{\lambda}_i, \boldsymbol{\delta}_i)||p_\Theta(\boldsymbol{z}_i|\boldsymbol{\lambda}_i, \boldsymbol{\delta}_i)) + \mathcal{L}(\Theta, \Phi; \boldsymbol{\lambda}_i, \boldsymbol{\delta}_i), \quad (5)$$

$$\mathcal{L}(\Theta, \Phi; \boldsymbol{\lambda}_i, \boldsymbol{\delta}_i) = \mathbb{E}_{q_\Phi(\boldsymbol{z}_i|\boldsymbol{\lambda}_i, \boldsymbol{\delta}_i)} \log p_\Theta(\boldsymbol{\lambda}_i, \boldsymbol{\delta}_i|\boldsymbol{z}_i) - KL(q_\Phi(\boldsymbol{z}_i|\boldsymbol{\lambda}_i, \boldsymbol{\delta}_i)||p_\Theta(\boldsymbol{z}_i)). \quad (6)$$

MVAE minimizes the first term from Equation 5 maximizing the second term (Equation 6), in which the first term is the neural network reconstruction loss and the second is the KL loss from $q_\Phi(\boldsymbol{z}_i|\boldsymbol{\lambda}_i, \boldsymbol{\delta}_i)$ and $p_\Theta(\boldsymbol{z}_i)$ (prior knowledge from distributed model). This study replaces the term $p_\Theta(\boldsymbol{z}_i)$ with the Gaussian distribution $\mathcal{N}(\boldsymbol{z}_i; 0, 1)$.

After representing the texts in a multimodal, robust, and appropriate way for OCL, the study can uses OCL algorithms to detect texts of interest. In OCL, the user defines an interesting class, and the OCL algorithm learns a classification model considering only documents of the interest class. Thus, the algorithm classifies a new document belonging to the interest class or not. Recent studies indicate that OCL is a competitive classification strategy with the advantage of reducing the user labeling effort [Alam et al. 2020, Tax 2001].

The study defines an OCL text classifier as a function $g : \mathcal{D} \rightarrow \mathcal{Y}$ that maps a textual document $\boldsymbol{d}_i \in \mathcal{D}$, with $D \in \mathbb{R}^n$, for a value $y_i \in \mathcal{Y}$, indicating how close document $\boldsymbol{d}_i$ is to belonging to the interest class. Thus, OCL aims to learn a function g^* from a training set $\mathcal{H} = \{\boldsymbol{d}_1, \boldsymbol{d}_2, \cdots, \boldsymbol{d}_n\}$ approximating the unknown mapping function g. Then, the classification through OCL is done by comparing each y_i with a threshold.

Among the OCL algorithms in the literature, the study chose the One-Class Support Vector Machines (OCSVM) [Tax and Duin 2004] since it achieves good performance when the user represents the instances appropriately, is considered one of the state-of-the-art in the area of OCL and use in different application and scenarios [Alam et al. 2020]. The OCSVM from [Tax and Duin 2004] aims to generate the smallest hypersphere (given a radius and center) that involves the examples of the interest class. Examples allocated to the edge of the hypersphere are the support vectors. According to [Tax and Duin 2004], the objective of OCSVM is to find a decision function capable of involving the textual documents of the interest class. Thus, the OCSVM wants to find the minimum volume hypersphere involving the interest documents according to Equation 7.

$$\boldsymbol{\mu}_{(c)} = \arg\min_{\boldsymbol{\mu} \in U} \max_{1 \leq i \leq m} \|\varphi(\boldsymbol{d}_i) - \boldsymbol{\mu}\|^2, \tag{7}$$

in which $\boldsymbol{\mu} \in U$ is a possible center in the feature space U associated with the function kernel φ, $\varphi(\boldsymbol{d}_i)$ maps $\boldsymbol{d}_i$ into a feature space defined according to the kernel, and $\boldsymbol{\mu}_{(c)}$ is the hypersphere center in which the highest distance between $\varphi(\boldsymbol{d}_i)$ to $\boldsymbol{\mu}_{(c)}$ is minimal.

OCSVM classifies a document as belonging to the interest class if its distance from the center is less than the radius r of the hypersphere ($y_i = dist(\varphi(\boldsymbol{d}_i), \boldsymbol{\mu}_{(c)})$, and $threshold = r$). For the hypersphere is not too large so that the false positive rate does not increase, the study adds a regularizer ν to accept a certain level of violation of the hypersphere decision function. Thus, the study desires to minimize the square of the hypersphere radius and the number of violations. The minimization function is given by Equations 8 and 9, in which ε_{d_i} is the external distance between $\varphi(d_i)$ and the surface of the hypersphere, and $\nu \in (0, 1]$ defines the smoothness level of the hypersphere volume.

$$\min_{\mu, \varphi, r} \quad r^2 + \frac{1}{m} \sum_{i=1}^{m} \frac{\varepsilon_{\boldsymbol{d}_i}}{\nu}, \tag{8}$$

$$\|\varphi(\boldsymbol{d}_i) - \boldsymbol{\mu}_{(c)}\|^2 \leq r^2 + \varepsilon_{\boldsymbol{d}_i} \quad \forall i = 1, ..., m. \tag{9}$$

4. Learning Textual Representations from Multiple Modalities to Detect Fake News Through One-Class Learning

In the experimental evaluation, the study proposes to compare the MVAE-FakeNews (MVAE-FK) with nine other unimodal and multimodal representation methods. Our goal is to demonstrate that the representations generated by MVAE-FK outperform others commonly used in the literature for news classification. The study uses three fake news collections. Furthermore, in order to evaluate the approach, this work proposes an adaptation of

the procedure k-Fold Cross-Validation, considering the OCL classification scenario with less labeling, in which the study uses 30%, 50%, 70%, and 100% of one fold to train (3, 5, 7, and 10% of the fake news) and nine folds to test (more details in dissertation).

Table 1 presents the results for the three textual collections considering each percentage of fake news used in training. The results compare the ten representation models of fake news. The proposed MVAE-FK obtained the highest F_1-Score in ten of the twelve evaluated scenarios. In the remaining two scenarios, the density information got the highest F_1. BoW with the TFIDF term weight obtained the lowest F_1 in all scenarios.

Table 1. Higher values of F_1-Scores of each representation technique considering the training percentage scenarios. The number of fake news equivalent to a percentage appears next to the %.

	%	TFIDF	TF	Binary	DBERTML	Density	LIWC	AE	VAE	MVAE-LIWC	MVAE-FK
Fake Br	3% (108)	0.029±0.01	0.600±0.01	0.617±0.00	0.574±0.01	0.621±0.03	0.597±0.01	0.622±0.01	0.632±0.01	0.637±0.00	**0.642±0.00**
	5% (180)	0.102±0.01	0.603±0.01	0.619±0.00	0.602±0.01	0.633±0.02	0.607±0.01	0.635±0.00	0.637±0.00	0.636±0.01	**0.644±0.00**
	7% (252)	0.182±0.01	0.607±0.00	0.620±0.00	0.618±0.01	**0.647±0.02**	0.614±0.00	0.638±0.00	0.637±0.00	0.637±0.00	0.645±0.00
	10% (360)	0.263±0.01	0.610±0.00	0.621±0.00	0.628±0.01	**0.650±0.01**	0.620±0.00	0.640±0.00	0.638±0.00	0.639±0.00	0.646±0.00
FCN	3% (31)	0.001±0.00	0,556±0.02	0,591±0.02	0.426±0.06	0.487±0.07	0.557±0.03	0.375±0.07	0.741±0.04	0.726±0.03	**0.805±0.02**
	5% (52)	0.010±0.01	0,582±0.01	0,605±0.01	0.568±0.05	0.575±0.08	0.587±0.02	0.631±0.04	0.796±0.03	0.736±0.02	**0.813±0.03**
	7% (73)	0.037±0.01	0,591±0.01	0,610±0.01	0.640±0.03	0.584±0.06	0.600±0.02	0.697±0.01	0.801±0.01	0.749±0.03	**0.811±0.02**
	10% (104)	0.110±0.02	0,596±0.01	0,614±0.00	0.706±0.02	0.625±0.03	0.613±0.01	0.722±0.03	0.804±0.02	0.753±0.03	**0.808±0.02**
FNN	3% (51)	0.001±0.01	0,327±0.01	0,355±0.01	0.321±0.01	0.325±0.02	0.379±0.01	0.337±0.01	0.365±0.01	0.386±0.01	**0.395±0.03**
	5% (85)	0.026±0.01	0,344±0.01	0,360±0.01	0.345±0.01	0.345±0.04	0.382±0.00	0.353±0.01	0.367±0.01	0.388±0.01	**0.403±0.03**
	7% (120)	0.081±0.01	0,349±0.00	0,361±0.00	0.353±0.01	0.353±0.01	0.384±0.00	0.358±0.00	0.367±0.00	0.390±0.00	**0.397±0.01**
	10% (170)	0.169±0.01	0,353±0.00	0,362±0.00	0.363±0.01	0.357±0.02	0.386±0.00	0.362±0.00	0.367±0.00	**0.393±0.00**	**0.393±0.01**

The study performed Friedman's statistical test with Nemenyi's post-test and Wilcoxon's statistical test to compare the representation methods. In addition to obtaining the best average ranking, MVAE-FK got statistical differences from unimodal representations. The test is in the dissertation. For FNN and FCN collections, the MVAE-FK, when trained with only 3% of labeled fake news, got better F_1-Scores than the other nine methods when these consider 10% of labeled fake news. MVAE-FK outperforms all other methods, obtaining better results even with few labeled fake news. It is worth mentioning that, without considering the representation of density, that got the best results considering 7% and 10% in the Fake.Br, it is possible to observe the same behavior. More qualitative results, such as a comparison of 2D projections, are found in the dissertation.

5. Detecting Relevant App Reviews through Multimodal One-Class Learning

In this experimental evaluation, the study proposes to compare our two multimodal representation methods, MVAE and MAE, with eight other representation methods from the literature. The study wants to demonstrate that the representations generated by MVAE and MAE outperform others usually used in the literature for app review classification. The study used the three datasets created by [Stanik et al. 2019] in the experimental evaluation and the representation proposed by them, which the study calls Maalej. The study uses the same adaptation of procedure k-Fold Cross-Validation mentioned in Section 4.

Table 2 presents the highest values of F_1-Score obtained by ten app review representation techniques on the ARE app review collection. Bold values indicate that the method obtained the highest value in the column. The complete results are in the dissertation.

Table 2. Highest F_1-Scores from OCSVM for each representation technique on the ARE dataset. The table also presents the number of reviews.

	3% #76	5% #127	7% #178	10% #254	22.5% #570	45% #1,142	67.5% #1,712	90% #2,283
Tfidf	0.55±0.02	0.57±0.02	0.58±0.01	0.60±0.01	0.60±0.03	0.63±0.01	0.64±0.02	0.65±0.02
Tf	0.54±0.00	0.54±0.00	0.54±0.00	0.54±0.00	0.57±0.00	0.57±0.00	0.57±0.00	0.57±0.00
Binary	0.54±0.00	0.54±0.00	0.54±0.00	0.54±0.00	0.57±0.00	0.57±0.00	0.57±0.00	0.57±0.00
Maalej	0.66±0.02	0.67±0.01	0.67±0.01	0.67±0.01	0.63±0.01	0.64±0.01	0.64±0.01	0.64±0.00
DBERTML	0.67±0.02	0.68±0.02	0.68±0.01	0.67±0.01	0.68±0.01	0.68±0.01	0.68±0.00	0.68±0.00
Density	0.66±0.03	0.66±0.02	0.65±0.02	0.64±0.02	0.65±0.02	0.64±0.01	0.64±0.01	0.64±0.02
AE	0.64±0.02	0.64±0.01	0.64±0.02	0.65±0.02	0.66±0.01	0.65±0.01	0.65±0.01	0.65±0.01
VAE	0.68±0.01	0.69±0.02	0.69±0.01	0.67±0.01	0.67±0.01	0.66±0.02	0.66±0.02	0.66±0.01
MAE	0.70±0.01	0.71±0.01	0.72±0.02	0.72±0.01	0.73±0.02	**0.75±0.02**	**0.77±0.01**	**0.78±0.01**
MVAE	**0.72±0.03**	**0.75±0.02**	**0.74±0.03**	**0.74±0.01**	**0.74±0.01**	0.73±0.01	0.73±0.02	0.72±0.02

MVAE obtained better results since it obtained the highest F_1-Scores than the other methods. Considering MAE, the MVAE was better in the scenario with less labeling. Moreover, MVAE with only 76 (3%) relevant reviews obtained the highest F_1-Scores than Maalej and all unimodal methods with 2,283 (90%) relevant reviews. Also, with only 127 (5%) relevant reviews, MVAE obtained competitive results related to MAE with 1,142 (45%) relevant reviews considering the F_1-Score. The study also performed Friedman's statistical test with Nemenyi's post-test in these experiments. The tests are in the dissertation.

6. Triple-VAE: A Triple Variational Autoencoder to Represent Events in One-Class Event Detection

The study uses 10 event collections for this experimental evaluation. The study chose specific experimental setups to simulate a more suitable scenario closer to real-world applications. First, the study uses event dates to separate training and testing. Events with older dates are from the training set. Second, the study explored using a few labeled instances in the training set. Thus, the study explored using 60, 120, 180, and 2000 events in the training set. In the test, the study uses 4000 interest events. Also, the study randomly selected 4000 events from different event datasets and added them to the test set.

The complete results are in the dissertation. However, for this manuscript, the study performed an analysis in relation to the Triple-VAE using 60 events compared to the other methods using 2000. This comparison is shown in Table 3. Triple-VAE achieved a higher F_1-Score than all other methods in 7 of 10 datasets. Furthermore, Triple-VAE achieved a higher F_1 in 3 collections than all other methods except DistilBERT. Thus, considering the scenario with fewer labeled events closer to real-world applications, Triple-VAE was the best method for event detection as it achieved the best F_1-Scores.

These results show that Triple-VAE was better than the other methods in learning highly non-linear relationships, redundancies, and dependencies between modalities, structuring the events in a dimensional space more suitable for OCL methods. Thus, our proposal structures events with more representativeness of their modalities in relation to

Table 3. Results in ten datasets considering the F_1. Values of TripleVAE are considering 60 events to train and the other representations methods 2000.

Datasets	Unimodal			Bimodal (DistilBERT — Lat-Long)				Trimodal			
	DBERTML	Lat-Long	Density	Concat	AE	VAE	BiVAE	Concat	AE	VAE	TripleVAE
War	**0,855**	0,665	0,732	0,677	0,685	0,688	0,726	0,681	0,688	0,689	0,780
Tsunami	**0,933**	0,647	0,680	0,667	0,684	0,675	0,811	0,670	0,685	0,680	0,918
Covid	**0,955**	0,677	0,748	0,732	0,771	0,771	0,949	0,737	0,773	0,775	0,946
Corruption	0,931	0,676	0,665	0,692	0,691	0,684	0,868	0,692	0,692	0,691	**0,958**
Earthquake	0,912	0,660	0,693	0,666	0,667	0,675	0,816	0,671	0,667	0,679	**0,916**
Immigration	0,928	0,668	0,664	0,693	0,762	0,780	0,900	0,693	0,781	0,814	**0,950**
Racism	0,940	0,666	0,825	0,688	0,729	0,745	0,910	0,694	0,754	0,786	**0,964**
Inflation	0,950	0,666	0,796	0,681	0,660	0,658	0,862	0,681	0,659	0,662	**0,953**
Terrorism	0,925	0,676	0,690	0,679	0,681	0,682	0,895	0,683	0,681	0,682	**0,937**
Agriculture	0,914	0,657	0,677	0,668	0,730	0,728	0,866	0,670	0,728	0,728	**0,979**

the other three methods. The study performed Friedman's statistical test with Nemenyi's post-test, considering all metric scenarios and datasets. The tests are in the dissertation.

7. Conclusions

This study presents a multimodal method developed to represent textual data considering the scenario of ATC through OCL. The method: (i) allows the use of a different number of modalities as input; (ii) allows the use of modalities with different dimensions; and (iii) is language and domain-independent. The study applies the multimodal method proposed in three real-world application domains: (i) fake news classification; (ii) relevant app reviews detection; and (iii) web sensing from news events. It is noteworthy that the study carried out an extensive empirical evaluation considering: (i) several multimodal variational autoencoder architectures; (ii) textual languages; and (iii) different sizes of training sets. The study highlights the following innovations and contributions with the development of this study:

Multimodal method to represent the texts in automatic text classification through OCL: the manuscript author proposes and develops a new multimodal method called Multimodal Variational Autoencoder (MVAE) and explores as modalities: (i) pre-trained embeddings from the DistilBERT multilingual; (ii) topic information from the high-density regions of the interest class; (iii) linguistic features of the texts; and (iv) geolocation (latitude and longitude). Any study can extend the proposed method to use more than three modalities.

Detecting fake news, relevant reviews, and interest events through proposed multimodal representations: the study proposes the MVAE to represent these texts for classification. The study explores three modalities for fake news: DistilBERT, density, and linguistic features. For relevant reviews, the study explores two modalities: DistilBERT and density. Finally, the study explores three modalities for interest event sense: DistilBERT, density, and geolocation (latitude and longitude). The study highlights satisfactory performance, outperforming other state-of-the-art methods in most scenarios.

Textual collection involving news events for web sensing tasks: In the study [Gôlo et al. 2021c], the study collects 183 textual datasets for the OCL. Each textual dataset has 6000 texts from the event titles of the Global Data of Events, Language, and Tone project. The study creates an OCTCMG library in a public repository.

Source code of the proposed MVAEs for the different applications investigated in this study: all the source codes developed in this study to pre-process collections, generation of representations, and OCSVM are available to the community in the repositories: Fake News, Relevant App Reviews, and Events. All source codes and the OCTCMG library are at `https://github.com/GoloMarcos/`.

As Limitation and future work, the study has: (i) The study uses the BERT model that limits the number of words used to generate the representation. Therefore, a future direction is to use language models that consider all the words in the text; (ii) This work used the k-Means algorithm to generate the density information. For future work, the study suggests other promising clustering algorithms; (iii) In the MVAE, there was no variation in some parameters due to time constraints. Therefore, a future direction is to variate these parameters for the MVAE to generate more robust models; (iv) Due to time constraints and the project's main focus (representation learning), the study uses only the OCSVM algorithm. However, one direction for future work is using other OCL algorithms; (v) investigating semi-supervised learning for OCL. Therefore, it is possible to combine the advantages of MVAE presented in this study and semi-supervised OCL; (vi) investigate graph modeling for the texts with multimodality in the OCL scenario; and (vii) Finally, this work explored the ATC through OCL using two separate steps: text representation and classification. Future work would be to classify and represent texts through OCL end-to-end, i.e., with a single learning process.

References

Aggarwal, C. (2018). *Machine Learning for Text*. Springer Publishing Company.

Alam, S., Sonbhadra, S. K., Agarwal, S., and Nagabhushan, P. (2020). One-class support vector classifiers: A survey. *Knowledge-Based Systems*, 196:1–19.

Cichosz, P. (2020). Unsupervised modeling anomaly detection in discussion forums posts using global vectors for text representation. *Natural Language Engineering*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of the 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minnesota. Association for Computational Linguistics.

Gao, J., Li, P., Chen, Z., and Zhang, J. (2020). A survey on deep learning for multimodal data fusion. *Neural Computation*, 32(5):829–864.

Gôlo, M., Caravanti, M., Rossi, R., Rezende, S., Nogueira, B., and Marcacini, R. (2021a). Learning textual representations from multiple modalities to detect fake news through one-class learning. In *Proc. of the Brazilian Symposium on Multimedia and the Web*.

Gôlo, M., Marcacini, R., and Rossi, R. (2019). An extensive empirical evaluation of preprocessing techniques and supervised one class learning algorithms for text classification. In *Proc. of the National Meeting on Artificial and Computational Intelligence*.

Gôlo, M. P., Araújo, A. F., Rossi, R. G., and Marcacini, R. M. (2022). Detecting relevant app reviews for software evolution and maintenance through multimodal one-class learning. *Information and Software Technology*, 151:106998.

Gôlo, M. P., Rossi, R. G., and Marcacini, R. M. (2021b). Triple-vae: A triple variational autoencoder to represent events in one-class event detection. In *Proceeding of the 2021 National Meeting on Artificial and Computational Intelligence.*, pages 643–654. SBC.

Gôlo, M. P. S., Rossi, R. G., and Marcacini, R. M. (2021c). Learning to sense from events via semantic variational autoencoder. *Plos one*, 16(12):e0260701.

Guo, W., Wang, J., and Wang, S. (2019). Deep multimodal representation learning: A survey. *IEEE Access*, 7:63373–63394.

Junior, D. and Rossi, R. (2017). Classificaçao automática de textos utilizando aprendizado supervisionado baseado em uma unica classe. *TCC em Sistemas de Informação*.

Katsaggelos, A. K., Bahaadini, S., and Molina, R. (2015). Audiovisual fusion: Challenges and new approaches. *IEEE*, 103(9):1635–1653.

Kumar, B. and Ravi, V. (2017a). One-class text document classification with OCSVM and LSI. In *Art. Intel. & Evolutionary Computations in Eng. Systems*. Springer.

Kumar, B. S. and Ravi, V. (2017b). Text document classification with PCA and one-class SVM. In *Proc. Int. Conf. on Frontiers in Intel. Computing: Theory and Applications*.

Li, Y., Yang, M., and Zhang, Z. (2018). A survey of multi-view representation learning. *IEEE transactions on knowledge and data engineering*, 31(10):1863–1883.

Manevitz, L. and Yousef, M. (2007). One-class document classification via neural networks. *Neurocomputing*, 70(7-9):1466–1481.

Manevitz, L. M. and Yousef, M. (2001). One-class svms for document classification. *Journal of machine Learning research*, 2(Dec):139–154.

Mayaluru, H. K. R. (2020). *One Class Text Classification using an Ensemble of Classifiers*. PhD thesis, Rheinische Friedrich-Wilhelms-Universität Bonn.

Rousseeuw, P. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.

Ruff, L., Zemlyanskiy, Y., Vandermeulen, R., Schnake, T., and Kloft, M. (2019). Self-attentive, multi-context one-class classification for unsupervised anomaly detection on text. In *Proc. of the Meeting of the Association for Computational Linguistics*.

Stanik, C., Haering, M., and Maalej, W. (2019). Classifying multilingual user feedback using traditional machine learning and deep learning. In *Int. Conf. Requirements Engineering*.

Tax, D. and Duin, R. (2004). Support vector data description. *Machine Learning*.

Tax, D. M. J. (2001). *One-class classification: Concept learning in the absence of counter-examples*. PhD thesis, Technische Universiteit Delft.

Wang, H., Bah, M. J., and Hammad, M. (2019). Progress in outlier detection techniques: A survey. *IEEE Access*, 7:107964–108000.

Xu, J. and Durrett, G. (2018). Spherical latent spaces for stable variational autoencoders. In *Proc. of the Conf. on Empirical Methods in NLP*. ACL.

Zhou, H., Yin, H., Zheng, H., and Li, Y. (2020). A survey on multi-modal social event detection. *Knowledge-Based Systems*, 195:105695.