*Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 3542–3550
Marseille, 11–16 May 2020

# Automatic In-the-wild Dataset Annotation with Deep Generalized Multiple Instance Learning

**Joana Correia [1,2], Bhiksha Raj [1], Isabel Trancoso [2]**
[1]Language Technologies Institute - Carnegie Mellon University, [2]INESC-ID - University of Lisbon
5000 Forbes Avenue, Pittsburgh, PA 15213, USA, Rua Alves Redol, 9, 1000-029 Lisboa, Portugal
{joanac, bhiksha}@cs.cmu.edu, isabel.trancoso@inesc-id.pt

### Abstract

The automation of the diagnosis and monitoring of speech affecting diseases in real life situations, such as Depression or Parkinson's disease, depends on the existence of rich and large datasets that resemble real life conditions, such as those collected from in-the-wild multimedia repositories like YouTube. However, the cost of manually labeling these large datasets can be prohibitive. In this work, we propose to overcome this problem by automating the annotation process, without any requirements for human intervention. We formulate the annotation problem as a Multiple Instance Learning (MIL) problem, and propose a novel solution that is based on end-to-end differentiable neural networks. Our solution has the additional advantage of generalizing the MIL framework to more scenarios where the data is still organized in bags, but does not meet the MIL bag label conditions. We demonstrate the performance of the proposed method in labeling an in-the-Wild Speech Medical (WSM) Corpus, using simple textual cues extracted from videos and their metadata. Furthermore we show what is the contribution of each type of textual cues for the final model performance, as well as study the influence of the size of the bags of instances in determining the difficulty of the learning problem.

**Keywords:** multiple instance learning, deep learning, automatic dataset annotation, in-the-wild data

## 1. Introduction

Speech is a complex bio-signal that is intrinsically related to human physiology and cognition. It has the potential to provide a rich bio-marker for health, and to allow a non-invasive route to early diagnosis and monitoring of a range of conditions and diseases that affect speech (Orozco-Arroyave et al., 2015)(Schuller et al., 2017), from depression, to Parkinson's disease, Alzheimer's disease, or a simple case of flu.

With the rise of speech related machine learning applications over the last decade, there has been a growing interest in developing tools that automatically perform non-invasive disease diagnosis and monitoring based on speech (Dibazar et al., 2002)(López-de Ipiña et al., 2013)(Lopez-de Ipiña et al., 2015)(Cummins et al., 2015)(Correia et al., 2016b)(Correia et al., 2018a).

The advantages of such tools are clear: their non-invasive nature allows for an increase in patient comfort; additionally, their automatic nature makes these solutions easy to scale.

However, one of the major factors that is slowing down the development of reliable tools for automatic diagnosis based on speech is the limited amount of speech medical data that is currently available, cataloged, and labeled.

The existing datasets are relatively small, and have been collected in very controlled conditions, showing limited applicability in real life situations. In fact, it has been shown that state-of-the-art models trained on these data can't maintain the performance reported when tested on clean data, when tested with data collected in-the-wild, where the noise and channel conditions are not controlled (Correia et al., 2018b).

An alternative is to use large and rich datasets collected from in-the-wild sources that better resemble real life situations. However manually labeling these datasets can be prohibitively expensive.

In this work we propose a novel approach to overcome this problem, and to automatically create and annotate datasets of in-the-wild speech medical data, of arbitrary size, with no human supervision, and no requirements for manual annotation. We propose using large scale, multimodal repositories, such as YouTube, as a source of in-the-wild speech medical data.

Our solution to automate the data gathering and labeling problem, is based on adding an extra filtering layer on top of existing retrieval algorithms on large-scale multimedia repositories, so as to significantly refine the search results. We formulate this solution as a generalized version of the multiple instance learning (MIL) problem, and propose solving it via neural networks. In the MIL scenario, a weakly supervised learning scenario, instances (or examples) are naturally organized into bags, for which a single label is assigned. A bag is assigned a positive label, if at least one of the instances in the bag is positive, otherwise the bag is assigned a negative label. Then bags, instances, and bag labels are used to learn instance labels. In the context of our problem, we have sets of search results associated with the search terms that were used to retrieve them from the multimedia repositories, obtained via whatever retrieval algorithm is implemented. These can be seen as bags of instances, for which, without further manual annotation we do not know which ones are relevant results and and which ones are not. What we know however, is that if we search for a relevant search term, the probability of finding at least one relevant search result is extremely high, and conversely, by searching for an unrelated search term, it is virtually guaranteed that no search results will be relevant. This scenario fits precisely with the MIL scenario. We further generalize the MIL scenario in the context of this problem by proposing a solution that allows the user to specify how many instances, or which fraction of the instances in the bag, need to be positive for the bag to be assigned a positive label.

Our proposed approach uses exclusively textual cues from the videos, namely its transcription, along with some of the available metadata, in order to perform the analysis of the video. We argue that this approach can later be extended to include acoustic and visual cues to further enrich the representation of the video examples, and improve the process of the dataset creation, but this option is not covered in this work.

We test the proposed approach with arguably two of the most important examples of diseases that affect speech, based on their their lifetime incidence rate are Depression and Parkinson's disease (PD). Specifically, depression is the leading cause of disability worldwide, with an increasing global prevalence of depression and depressive symptoms in recent decades (Vos et al., 2016). The lifetime prevalence of depression ranges from 20% to 25% in women, and 7% to 12% in men (Organization, 2002). On the other hand, PD is the second most common neurological problem in the elderly, after Alzheimer's Disease, affecting 1 to 2 in every 100 persons over 60 years old (De Lau and Breteler, 2006). Out of all the neurological disorders, PD is the fastest growing one (Muangpaisan et al., 2011).

The contributions of this work are twofold:

- A deep learning based solution for the generalized formulation of MIL framework, where instead of determining the positiveness of a bag from a single positive instance, it is possible to specify how many or what fraction of the bag instances need to be positive to assign a positive bag label;
- Using the proposed solution to automate the creation of speech and language resources from in-the-wild multimedia repositories in the context of speech affecting diseases, without any requirements for manual annotation.

This paper is organized as follows: In Section 2. we review the related work, namely the applications of MIL in several problems and some of the proposed variations and generalizations of MIL; In Section 3. we describe MIL in further detail, along with the proposed generalization and our deep learning based solution; Section 4. describes the collection and manual annotation process of the in-the-Wild Speech Medical (WSM) Corpus, an in-the-wild dataset, collected from YouTube, for depression and PD, a first of its kind dataset that features real life noise and channel conditions. These conditions are not controlled for, and not considered a factor when collecting and annotating the data. In Section 5. we describe the experiments performed in this work which focus on using the proposed deep generalized MIL solution to automatically annotate the WSM corpus. Finally, in Section 6. we draw some conclusions and discuss future work.

## 2. Related work

Multiple instance learning (Dietterich et al., 1997)(Maron and Lozano-Pérez, 1998) is a special case of weakly supervised learning where instances (or examples) are naturally organized into bags, for which a single label is assigned. A bag that contains at least a positive instance is assigned a positive label, and a bag that only contains negative instances is assigned a negative label. At training time the instance labels are not available, and at test time, the task is one or both of the following, depending on the application: inferring bag labels, and/or inferring instance labels.

MIL has been used in several contexts such as medical imaging segmentation (Quellec et al., 2017)(Kraus et al., 2016)(Ilse et al., 2018) where an image is typically described by a single label, but the region of interest is not given. In this case, the bag is the image and each segment of the image is an instance. Other examples include drug activity prediction (Dietterich et al., 1997), image annotation and retrieval (Carneiro et al., 2007), text categorization (Liu et al., 2018), and object detection (Zhang et al., 2006), among others.

There are multiple approaches to address the MIL problem. Arguably, one of the early, most popular ones is (Andrews et al., 2003)'s solution, that describes two algorithms based on SVM's that formulate the MIL problem as a maximum-margin problem that can be solved via mixed integer quadratic programs: mi-SVM and MI-SVM.

More recent works have proposed solutions for the MIL problem via deep neural networks, as is the case of the pioneering work of (Wu et al., 2015). Others, such as (Ilse et al., 2018), proposed a formulation for the MIL problem as learning the Bernoulli distribution of the bag labels, which is parametrized by a neural networked with an attention mechanism.

A few other works have tried to solve generalized versions of the MIL problem, usually by changing the bag label assumptions, or by acknowledging label noise. In (Correia et al., 2016a), the authors propose a generalization of mi-SVM and MI-SVM: $\theta-$mi-SVM and $\theta$-MI-SVM, where the bag labels are determined not by the presence of at least one positive instance in the bag, but instead by the presence of a fraction $theta$ of positive instances out of all the instances in the bag. This solution is based on maximum-margin algorithms, and can be applied in scenarios where one positive instance is not enough evidence for a positive bag. In this case, it was applied in the context of inferring the polarity of movie reviews. Other such works include (Li and Vasconcelos, 2015), where the authors tackle the problem of label bag noise in the context of semantic image retrieval, by introducing the notion of soft bags, i.e. bags that can contain both positive and negative instances, regardless of their label. This allows for negative bags to have some positive instances in them, which are considered to be noise. The problem is then solved via a large-margin algorithm.

## 3. Methodology

### 3.1. Multiple Instance Learning

One of the simplest learning scenarios is fully supervised binary classification, i.e. a scenario where the training data are pairs of instances and labels, $X = \{\{x_1, y_1\}, \{x_2, y_2\}, ..., \{x_n, y_n\}\}$, where $x_i \in^{\mathbf{D}}$ is the $i^{th}$ training instance, and $y_i \in \{0, 1\}$ is its corresponding binary label. Then $X$ is used to train a model such that, for a new instance $x$, it is possible to estimate its label $\hat{y} = f(x)$, where $f(.)$ is the function that maps instances to labels.

However, learning in a fully supervised scenario is not always possible. This is frequently the case of applications that rely on datasets that are collected from in-the-wild sources, where the data is unstructured, or loosely structured, but unlabeled. In these cases, manual annotation can quickly become too expensive, either because of the massive number of examples that have to be annotated, or because each example is expensive to annotate. So other solutions have to be adopted.

The MIL scenario arises as a weakly supervised learning alternative when fully supervised learning is impossible, but there is still some structure to the data. In this case, the assumption is that the instances, or examples, are organized into permutation-invariant bags $B = \{x_1, x_2, ..., x_k\}$, where each $x \in^{\mathbf{D}}$ is an instance, and $k$ is the size of the bag, which can vary for each bag $B$. Again, we emphasise that the instances are independent and that there is no specific order for them in the bag. Each bag is associated with a label $Y \in \{0, 1\}$, such that the training dataset can be represented by $X = \{\{B_1, Y_1\}, \{B_2, Y_2\}, ..., \{B_n, Y_n\}$, where $n$ is the number of bags in the dataset. In MIL, during training, the instance level labels for the instances of any given bag $B$, $\{y_1, y_2, ..., y_k\}$, are not available, only the bag level label $Y$ is.

The bag level label $Y$ for any given bag $B$ is defined as:

$$Y = \begin{cases} 1, & \text{if } \sum_{i=0}^{k} y_i \geq 1 \text{ ;} \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

Or more compactly as:

$$Y = \max_i (y_i) \tag{2}$$

This bag level label definition can be interpreted as "if there is at least one positive example in the bag, the bag label is positive". In a way, positive instances can be seen as the key that opens a lock in a set of keys, which is the bag. During training, the only information that is available is if it is possible to open the lock with a given set of keys, but not with which key specifically.

The goals in the MIL scenario are twofold, given a new bag $B$ of size $k$: learning to predict the bag labels $Y$, and learning to predict the instance labels $y_1, ...y_k$.

### 3.2. Generalized Multiple Instance Learning

There are situations however, where the presence of one positive instance is not enough evidence to justify assigning a positive label to a bag. In such scenarios, the traditional MIL framework falls short. However, this framework can be reformulated in more general terms, such that it can be used to solve problems as the above mentioned.

Let us assume again that a bag $B = \{x_1, x_2, ..., x_k\}$ is a collection of instances $x_i \in^{\mathbf{D}}$. In the generalized MIL scenario, the bag label $Y$ for any given bag $B$ is defined as:

$$Y = \begin{cases} 1, & \text{if } \sum_{i=0}^{k} y_i \geq k\delta \text{ ;} \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

where $\delta \in [0, 1]$ is a parameter that determines what is the minimum fraction of the instances in the bag that is required to have a positive label, for the bag label to be positive. It follows that, in this scenario, even negative bags can have some instances with positive labels so long as they verify $\sum_{i=0}^{k} i_m < k\delta$.

We note that the generalized MIL framework corresponds to the conventional MIL framework when $\delta = \frac{1}{k}$.

However, solving the generalized MIL problem is more complex the MIL one, since the trick of computing the maximum of the instance labels for a given bag, as stated in Eq. 2, can not be used for values of $\delta$ greater than $\frac{1}{k}$.

### 3.3. Deep Generalized Multiple Instance Learning

Considering the success that deep learning approaches have had over the course of the last decade, it is only natural that we adopt them over the traditional SVM based approaches previously used to solve the MIL problem. They have the additional advantage of allowing flexible training strategies, since they can be trained end-to-end via backpropagation, so long as the strategy to pool the instance labels into bag labels is differentiable.

Let us assume for the sake of simplicity that instances are represented by features, that are obtained via arbitrary transformations, parametrized by neural networks, such that $h = f_\psi(x)$, where $h$ is the hidden representation of the instance $x$. $h$ is obtained after performing the transformation $f_\psi(.)$, where $\psi$ are the transformation parameters.

Given a bag of hidden representations of instances, and its respective label $B = \{\{h_1, h_2, ..., h_k\}, Y\}$, the proposed approach is to define two more fully differentiable transformations that convert the hidden representation of the instances $h_i$, to instance labels $y_i$: $y = g_{1\phi}(h_i)$, where $y \in 0, 1$, and $\phi$ are the transformation parameters; and another transformation that pools the instance labels $\{y_1, y_2, ..., y_k\}$ into a bag label $Y$, following the constrains of the generalized MIL framework, as stated in Eq. 3: $Y = g_{2\theta}(y_1, y_2, ..., y_k)$, where $\theta$ are the transformation parameters.

The first of the two transformations, $g_1(.)$, is trivial and any multi-layer perceptron (MLP) can be trained to parameterize it, with the constraint that the output layer contains only one unit and a sigmoid activation function.

For the second transformation, we propose a pooling scheme that guarantees the constraints of not only the MIL framework, but also the generalized MIL framework:

$$Y = \frac{e^{\sum_{i=0}^{k} y_i - k\delta}}{e^{\sum_{i=0}^{k} y_i - k\delta} + 1}, \tag{4}$$

This transformation is essentially the sigmoid function for the sum of the labels of the bag instance, with a horizontal skew of $k\delta$. The parameter $\delta \in [0, 1]$ determines the fraction of the instances in the bag that have to be positive to assign the bag a positive label. When $\theta = \frac{1}{k}$, the horizontal skew disappears, and the sigmoid of the sum of the instance labels can be interpreted as a differentiable approximation of the maximum operator. For other feasible values of $\theta$, the horizontal skew on the sigmoid of the sum of the instance labels, will ensure that the function output, i.e., the bag label, is zero when the sum of the instance labels is smaller than $k\delta$, i.e., the number of positive instances in the bag is smaller that the minimum number of instances necessary to assign a positive label to the bag.

In order to train a neural network with parameters $\theta$ that performs the transformation $g_{2\theta}(.)$, the loss to be minimized during training must reflect the generalized MIL constraints. A possible loss based on the binary crossentropy of the bag labels, as computed in Eq.4 is:

$$\begin{cases} L = -t\log(s) - (1-t)\log(1-s), \\ s = \frac{e^{\sum_{i=0}^{k} p_{y\,i} - k\delta}}{e^{\sum_{j=0}^{k} p_{y\,i} - k\delta} + 1} \end{cases} \tag{5}$$

where $t$ is the groundtruth bag label, and $s$ is the predicted score for a bag label. $s$ is computed by applying the formula stated in Eq. 4 to the individual scores of the predicted instance labels $p_{y\,j}$.

## 4. WSM Corpus

The WSM Corpus is an audio-visual corpus of videos collected from the multimodal repository YouTube, that feature videos related to several speech affecting diseases: Depression, Parkinson's disease and the common cold (Correia et al., 2018b). A first version of this corpus was collected in February of 2018, with videos published between January 2007 to February 2018, containing approximately 60 videos per speech affecting disease. More recently, the WSM has been expanded to approximately quadruple the size, and is planned to continue to grow, and to include additional speech affecting diseases.

The dataset was collected by using a combination of the official YouTube API and scrapping tools to retrieve a list of results for several queries related to each target disease, p.e., in the case of depression, “[*target disease*] vlog". For each result, the following information was collected: the video and audio; the metadata (including the title, description, channel and video identifiers, etc.); the video's transcription; and the comments to the video.

We note that the video's transcription is automatically generated by YouTube (only for videos in English), using a large scale, semi-supervised deep neural network for acoustic modeling (Liao et al., 2013), unless a manual transcription is provided by a user.

Each video in WSM Corpus was manually labeled for the presence of a subject affected by the target disease, so each video contains three binary labels for the presence of a subject affected by Depression, PD, or a cold, respectively.

We emphasise that, because the videos are collected from online repositories, and there is no opportunity for the subjects in the videos to be interviewed by medical professionals, we determine their health status solely based on their self assessment. I.e. if a subject in a video claims that he or she is currently affected by the target disease, e.g. depression, that is enough evidence to assign a positive label for that disease. While we recognise that not all the subjects may be sincere in their self-diagnosis, we assume that the amount of label noise is negligible, given that the subjects usually do not have a motivation to misrepresent their health status in the context of the videos.

In the context of this work, we will use only the depression and PD subsets of the WSM Corpus, which amount to 550 videos in total. Out of the 550 search results, 59 and 26 are positive for the presence of a subject self diagnosing with depression and Parkinson's disease respectively.

## 5. Experiments and results

In this section, we describe the experiments which focus on using the proposed deep generalized MIL solution to automatically label the videos of the WSM Corpus featuring subjects affected by Depression and PD.

We take advantage of the existing structure on the WSM Corpus, where the videos are associated to the search term that was used to retrieve it, as well as a time window for the upload date. This structure can easily be translated to the (generalized) MIL framework, where a bag is the set search results obtained for a given search term and time window. If the search term used to generate the bag relates to the target disease, and particularly, we believe that among the search results will be examples of videos featuring people affected by the target disease, we give the bag a positive label for that disease. As an example, let us say the target disease is depression and the search term is "depression vlog", then we assume that among the retrieved search results there will be some that are relevant. Otherwise, if the search term is not likely to generate relevant results, such as "depression lecture", we assign a negative label to the set of search results. We note that while the results are still related to depression in the case of the negative bag just described, the videos are very unlikely to contain subjects that are currently affected by depression, and are much more likely to contain healthy subjects, such as medical doctors, therapists, researchers or journalists, describing some aspect of depression, mostly from a third party perspective. We argue that generating such negative bags poses a much more interesting and nuanced problem than simply generating negative bags of videos with completely unrelated content. Any model that is successfully trained with such data is much more likely to be useful in real life situations to detect subjects affected by a target disease that if the control examples were completely unrelated. From a class separation perspective, we argue that these negative examples are much closer to the true class decision boundary than unrelated ones, which allows for a better estimation when training a model to learn it.

In regards to the information derived from the search results used in this work, we used exclusively information derived from the textual components of the video, and relay the study of acoustic and visual features for future work. In particular we used the transcription of the video, the title, the description, and the top 5 comments.

In regards to the feature extraction process used in this work, we briefly describe the adopted process based on Sentence-BERT (Reimers and Gurevych, 2019), that was used to encode the textual cues.

Then, we establish an upper bound for the performance that can be achieved on the task of automatically labeling the WSM corpus using the chosen features, by training a fully supervised model on the dataset with the manually obtained labels. After this, we compare the performance of the fully supervised model to the performance obtained of a similar model but in a MIL scenario, where the instance labels are unavailable during training. We also study the contribution of each type of textual cues for the performance of the final model: the transcription, the title and description of the videos, and the top n comments to the video. Finally we study the influence of the bag size in the performance of the models, and as a sanity check, show that when bags are smaller, the MIL problem is easier to solve, and that in the extreme case where the bag size is one, the problem becomes a fully supervised learning problem, since all the instance labels correspond to their respective bag label.

### 5.1. Feature extraction

BERT (Devlin et al., 2018) is considered the state-of-the-art in encoding language representations, and is based on bidirectional transformers. It is designed to generate representations from unlabeled text by jointly conditioning on both its left and right context. However, it is not optimized for long sentences or even full text documents. Sentence-BERT (SBERT)(Reimers and Gurevych, 2019), is a modification of the BERT network, using siamese and triplet networks, in order to derive meaningful sentence embedding of fixed sized, for arbitrarily sized sentences, converting them into feature vectors of 768 dimensions. With SBERT the similarity between sentences can be computed by any similarity measure such as cosine similarity.

As such we adopt a pre-trained version of SBERT that was first trained on Natural Language Inference (NLI) data, then fine-tuned on AllNLI, and on the semantic text similarity (STS) benchmark training set, obtaining an STS score of 85.29, as reported by the authors.

We use this pre-trained model to embed the three documents associated with a search result: 1) the transcription of the video; 2) the title and description of the video provided by the user; and 3) the top 5 comments on the video, sorted by popularity. Thus each search result is characterized by three 768-dimensional vectors. We repeat this for the complete dataset of 550 search results, which amounts to a total of 1650 embeddings.

Some of the videos did not have any comments, therefore a random 768-dimensional vector was generated, which represents a random sentence/document.

### 5.2. Fully supervised upper bound

As mentioned before, the WSM Corpus contains manual annotations. While it is not realistic to assume that these will be available in a real-life application, they are useful to perform a number of tasks, such as for evaluation purposes, for semi supervised learning tasks, etc. In this case, we use the manual annotations of the WSM corpus to perform the fully supervised learning task of predicting the presence of subjects affected by two diseases, depression and Parkinson's disease, from SBERT embeddings of the transcription, title and description, and top 5 comments of YouTube videos. This experiment is useful to determine what is the upper bound of the performance that could be obtained with the WSM, if all the instance labels were available. In other words, by comparing the following results with the performance of a similar model in a (generalized) MIL learning scenario, it is possible to quantify the loss of knowledge when the instance labels are not available and only the bag structure and bag labels of the dataset are.

The architecture of the fully supervised model is shown in Figure 1 (left). It comprises three streams of MLPs, one for each type of embedding of the three documents available, with 768 dimensions each. Each stream contains two fully connected layers, with 256 and 64 units, respectively, and both with ReLU activation, and a dropout rate of 0.2. The streams are fused by concatenating the three hidden representations. The network has two more fully connected layers, the first of them with 64 units and a ReLU activation and a dropout rate of 0.5, and finally, a 1 unit output layer with sigmoid activation.

The network was trained with a binary cross entropy loss, and RMSProp optimizer algorithm, over 60 epochs, with a learning rate of 0.001, and early stopping conditions based on the development loss.

The model was trained with 400 examples and tested against 150 examples for both Depression and Parkinson's disease. The performance of this model was measured in F1 score, since there is a significant class imbalance. The fully supervised model obtained an F1 score of 0.69 and 0.67 on the test set for Depression and PD, respectively. These results are also shown along with others in the bar charts on Figures 2 and 3 for Depression and Parkinson's disease, respectively, as the leftmost columns in the color red.

These results by themselves are not very meaningful, since the features or the model architecture, among others, could be changed in an attempt to improve the performance of the model, however this is the most fair upper bound to performance of the experiments in the following sections.

### 5.3. Deep Generalized MIL performance

The main contribution of this work experimentally verified in this section, where we test the proposed deep generalized MIL solution, in labeling the WSM Corpus, without access to any of the manual labels, and having only access to the bag structure and to the bag labels.

As mentioned in Section 4., the WSM has 11 bags of 50 examples each. We use 8 of these bags, which total 400 instances for training and, the remaining 3 for testing purposes, which contain 150 examples. The distribution of the train and test examples is the same as in the experiments reported in Section 5.2., for comparison purposes.

For each of the instances we compute the same SBERT embedding for the three available documents: the transcription, the title and description and the top 5 comments, and use these 3 768-dimensional vectors as the input to the network. The architecture of this network is similar to the one described in Section 5.2., where each instance is processed by a 3 stream network. These streams are then fused and used to generate an instance prediction. The key difference in this case, is that this process is repeated for all the instances in the bag, and the prediction of the bag label is made according to Eq. 4. Only then the loss is computed according to Eq. 5, and the weights updates, through backpropagation. A good strategy to implement this network it to set the batch size to be the bag size and process the instances if the same bag in sequence, so that they are all processed in the same batch. By doing so, the loss is accumulated over the whole bag and the predictions for all the instances in the bag are computed with the same network weights.

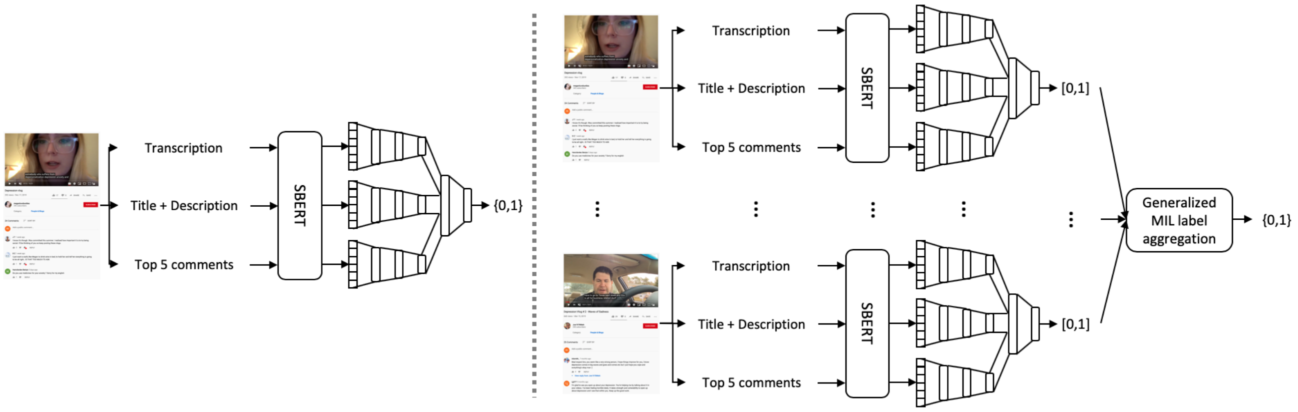

Figure 1: Left: architecture of the fully supervised model that estimates the upper bound of the performance that can be obtained in labeling the WSM, given the feature choice and model architecture. Right: architecture of the proposed deep generalized MIL solution.

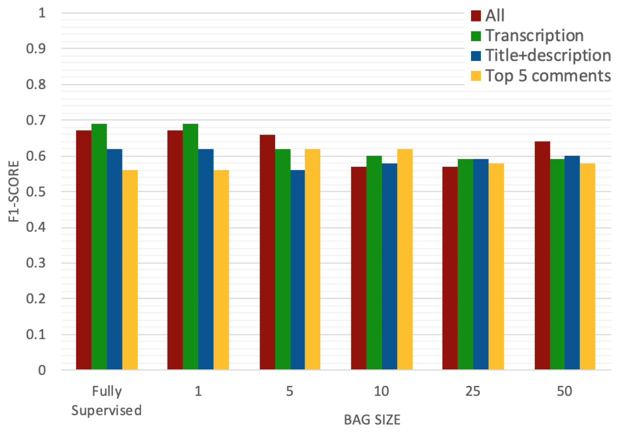

Figure 2: Summary of the performance in F1 score of all the models trained to estimate the Depression labels of the WSM Corpus, for different bag sizes and sources of textual cues.

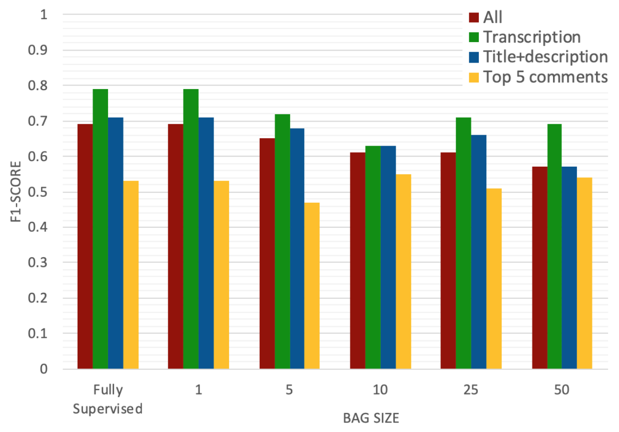

Figure 3: Summary of the performance in F1 score of all the models trained to estimate the Parkinson's disease labels of the WSM Corpus, for different bag sizes and sources of textual cues.

We summarize the architecture of the proposed deep generalized MIL network in Figure 1 (right).

The network was trained with similar parameters as the one presented in Section 5.2., except $\delta$ was set to $2/k$, where $k$ is the size of the bag.

The performance of this network on the test set at instance label level, was an F1-score of 0.57 and 0.64, for Depression and Parkinson's disease respectively. These results are also shown on Figures 2 and 3, on the set of columns above the label 50, with the color red. We also note that the network was able to achieve an error of zero at the bag level labels, however we consider that this result is not relevant in the scope of this work, which is instance level prediction, so we will not mention it in further experiments.

We note that for the case of Parkinson's disease, there is a significant drop in performance, while for depression the drop in performance is smaller. Nevertheless, in both situations we can experimentally confirm the hypothesis that learning in a (generalized) MIL is a harder problem than a fully supervised one.

### 5.4. Contribution of each type of document

So far, we have only shown experiments where we take advantage of the three types of documents available for each video, the transcription, the title and description, and the top 5 comments, at the same time. However, it a reasonable assumption that the contribution of each document could vary widely for the final instance label prediction. In this section, we study the contribution of each document for the performance of the proposed networks. We achieve this by making a minor modification to the network described in Section 5.3.: removing two of the three documents, and their respective streams in the network. Following this change, the only other necessary change was to remove the concatenation layer that merged the three streams into one vector.

We perform the same experiments with the same data partitions as described in section 5.3., for each one of the three types of documents. The performance in F1-score of the instance level label prediction is summarized in Table 5.4. for Depression and PD.

As can be seen for Parkinson's disease, the source of data

| Document\Disease | Depression | Parkinson's disease |
|---|---|---|
| **Transcription** | 0.59 | 0.69 |
| **Title + description** | 0.60 | 0.57 |
| **Top 5 comments** | 0.58 | 0.54 |

Table 1: Performance in F1 score of the proposed deep MIL network for one type of textual cue at a time, for Depression and Parkinson's disease

that contains the most useful information for this problem is the transcription, which on its own outperforms the networks trained on the three documents. Intuitively this result is not unexpected, since it is reasonable to assume that the content of the conversation by the subject of the video will be far more important to determine the health status than the title and description of the video, or the comments to the video. However this is not the case for depression, where the three models have a very similar performance.

We note however, that across the two diseases, the model trained only with the top 5 comments was the one with the poorest average performance. One of the reasons behind this may be that not all the videos have comments, and for these cases random embedding vectors were generated to replace them. The models trained on title and description of the video, while not so poor as the one trained on the top 5 comments, still performed worse that the model trained on the transcription.

### 5.5. Influence of bag size

Another variable in this problem that is interesting to study is the influence of the bag size. Intuitively, one would expect as the size of the bags get bigger, the learning problem becomes harder. In this section we test this hypothesis for bag sizes of 5, 10, and 25. Additionally we also perform the same experiments by setting the bag size to 1, where we expect to see the same results as the ones reported in Section 5.2., since the two problems become equivalent: the instance labels are completely determined by the bag labels.

We note that the experiments reported in Sections 5.3., and 5.4., were also repeated for different bag sizes: 1, 5, 10, and 25, so that we could make a complete report on the influence of these two variables: bag size, and input documents.

The networks were trained with the same parameters as before, as well as trained and tested in the same partitions of the data. However, depending on the bag size, each bag was randomly divided into smaller bags. The instances from the original bags were not mixed. The new bag labels were assigned based on the aggregation of the manually obtained instance level labels for that bag, following Eq. 3.

These results for this experiment, for both Depression and Parkinson's disease are summarized in Table 5.5.. Furthermore, the results are also shown in Figures 2, and 3, for Depression and Parkinson's respectively. In these two figures it is possible to compare the performance of all the models with different bag sizes and input data, and the fully supervised upper bound.

Again, in this set of experiments, we are able to confirm the hypothesis that, as a rule of thumb that larger bags are associated to a harder learning problem. This occurs because for bigger bags, the restrictions on the instance labels are smaller. In fact, in the extreme case of a positive bag with infinite samples, we would be almost in a completely unsupervised learning scenario, other that having the prior of knowing that a fraction of the instance belonged to the positive class.

| Bag size | Document\Disease | Depression | Parkinson's disease |
|---|---|---|---|
| **1** | **All** | 0.67 | 0.69 |
| | **Transcription** | 0.69 | 0.79 |
| | **Title + description** | 0.62 | 0.71 |
| | **Top 5 comments** | 0.56 | 0.53 |
| **5** | **All** | 0.66 | 0.65 |
| | **Transcription** | 0.62 | 0.72 |
| | **Title + description** | 0.56 | 0.68 |
| | **Top 5 comments** | 0.62 | 0.47 |
| **10** | **All** | 0.57 | 0.61 |
| | **Transcription** | 0.60 | 0.63 |
| | **Title + description** | 0.58 | 0.63 |
| | **Top 5 comments** | 0.62 | 0.55 |
| **25** | **All** | 0.57 | 0.61 |
| | **Transcription** | 0.59 | 0.71 |
| | **Title + description** | 0.60 | 0.66 |
| | **Top 5 comments** | 0.58 | 0.51 |

Table 2: Performance in F1 score for the proposed deep MIL network for diferent sizes of bags, and different types of textual cues, for Depression and Parkinson's disease.

## 6. Conclusion and Future Work

This work's motivation was to overcome the problem of the lack of existence of large and rich speech medical dataset with which to train deep and complex models for the detection and monitoring of speech affecting diseases. We proposed that a solution for this problem would be to mine the data from online multimedia repositories, such as YouTube, and automate the labeling process. More specifically, the solution that we proposed for the automation of the annotation process, included formulating this problem in the generalized MIL framework, for which we proposed a solution based on deep neural networks. To achieve this, our major contribution was to adopt a new loss function that verified the conditions of the generalized MIL framework and that the same time was fully differentiable to allow training via backpropagation. We tested the proposed framework on the WSM Corpus, specifically for the detection of subjects affected by Depression and Parkinson's disease in YouTube videos. We used features derived from the transcription, metadata, and comments of the video. In our experiments, we were able to confirm several interesting phenomena, namely: We confirmed that the bag size has influence in determining the difficulty of the learning problem - larger bags create harder learning problems; we quantified the contribution of each type of document used to describe the data, and concluded that, regardless of the bag size, the transcription consistently carried the most information to determine the heath status of the video's subject.

Finally, we argue that the core of this work lies on the new generalized MIL formulation, and that details such as features or input data, can be trivially changed so that our solution can be used in different domains, as long as the data still maintains the structure imposed by (generalized) MIL.

As for future work, we will plan to include multimodal sources of data to perform the dataset labeling, such as the audio of video. We also plan do further develop the proposed generalized MIL solution such that we can decrease that gap between its performance and the performance of a comparable model in a fully supervised scenario.

## 7. Acknowledgements

This work was supported by Portuguese national funds through Fundação para a Ciência e a Tecnologia (FCT) with reference UID/CEC/50021/2019.

## 8. References

Andrews, S., Tsochantaridis, I., and Hofmann, T. (2003). Support vector machines for multiple-instance learning. In Advances in neural information processing systems, pages 577–584.

Carneiro, G., Chan, A. B., Moreno, P. J., and Vasconcelos, N. (2007). Supervised learning of semantic classes for image annotation and retrieval. *IEEE transactions on pattern analysis and machine intelligence*, 29(3):394–410.

Correia, J., Trancoso, I., and Raj, B. (2016a). Adaptation of svm for mil for inferring the polarity of movies and movie reviews. In 2016 IEEE Spoken Language Technology Workshop (SLT), pages 258–264. IEEE.

Correia, J., Trancoso, I., and Raj, B. (2016b). Detecting psychological distress in adults through transcriptions of clinical interviews. In International Conference on Advances in Speech and Language Technologies for Iberian Languages, pages 162–171. Springer.

Correia, J., Raj, B., and Trancoso, I. (2018a). Querying depression vlogs. In 2018 IEEE Spoken Language Technology Workshop (SLT), pages 987–993. IEEE.

Correia, J., Raj, B., Trancoso, I., and Teixeira, F. (2018b). Mining multimodal repositories for speech affecting diseases.

Cummins, N., Scherer, S., Krajewski, J., Schnieder, S., Epps, J., and Quatieri, T. F. (2015). A review of depression and suicide risk assessment using speech analysis. *Speech Communication*, 71:10–49.

De Lau, L. M. and Breteler, M. M. (2006). Epidemiology of parkinson's disease. *The Lancet Neurology*, 5(6):525–535.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding.

Dibazar, A. A., Narayanan, S., and Berger, T. W. (2002). Feature analysis for automatic detection of pathological speech. In Engineering in Medicine and Biology, 2002. 24th Annual Conference and the Annual Fall Meeting of the Biomedical Engineering Society EMBS/BMES Conference, 2002. Proceedings of the Second Joint, volume 1, pages 182–183. IEEE.

Dietterich, T. G., Lathrop, R. H., and Lozano-Pérez, T. (1997). Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence*, 89(1-2):31–71.

Ilse, M., Tomczak, J. M., and Welling, M. (2018). Attention-based deep multiple instance learning. *arXiv preprint arXiv:1802.04712*.

Kraus, O. Z., Ba, J. L., and Frey, B. J. (2016). Classifying and segmenting microscopy images with deep multiple instance learning. *Bioinformatics*, 32(12):i52–i59.

Li, W. and Vasconcelos, N. (2015). Multiple instance learning for soft bags via top instances. In Proceedings of the ieee conference on computer vision and pattern recognition, pages 4277–4285.

Liao, H., McDermott, E., and Senior, A. (2013). Large scale deep neural network acoustic modeling with semi-supervised training data for youtube video transcription. In 2013 IEEE Workshop on Automatic Speech Recognition and Understanding, pages 368–373. IEEE.

Liu, B., Xiao, Y., and Hao, Z. (2018). A selective multiple instance transfer learning method for text categorization problems. *Knowledge-Based Systems*, 141:178–187.

López-de Ipiña, K., Alonso, J.-B., Travieso, C. M., Solé-Casals, J., Egiraun, H., Faundez-Zanuy, M., Ezeiza, A., Barroso, N., Ecay-Torres, M., Martinez-Lage, P., et al. (2013). On the selection of non-invasive methods based on speech analysis oriented to automatic alzheimer disease diagnosis. *Sensors*, 13(5):6730–6745.

Lopez-de Ipiña, K., Alonso, J. B., Solé-Casals, J., Barroso, N., Henriquez, P., Faundez-Zanuy, M., Travieso, C. M., Ecay-Torres, M., Martinez-Lage, P., and Eguiraun, H. (2015). On automatic diagnosis of alzheimer's disease based on spontaneous speech analysis and emotional temperature. *Cognitive Computation*, 7(1):44–55.

Maron, O. and Lozano-Pérez, T. (1998). A framework for multiple-instance learning. In Advances in neural information processing systems, pages 570–576.

Muangpaisan, W., Mathews, A., Hori, H., and Seidel, D. (2011). A systematic review of the worldwide prevalence and incidence of parkinson's disease. *Journal of the Medical Association of Thailand*, 94(6):749.

Organization, W. H. (2002). The world health report 2002: reducing risks, promoting healthy life. World Health Organization.

Orozco-Arroyave, J. R., Belalcazar-Bolanos, E. A., Arias-Londoño, J. D., Vargas-Bonilla, J. F., Skodda, S., Rusz, J., Daqrouq, K., Hönig, F., and Nöth, E. (2015). Characterization methods for the detection of multiple voice disorders: Neurological, functional, and laryngeal diseases. *IEEE journal of biomedical and health informatics*, 19(6):1820–1828.

Quellec, G., Cazuguel, G., Cochener, B., and Lamard, M. (2017). Multiple-instance learning for medical image and video analysis. *IEEE reviews in biomedical engineering*, 10:213–234.

Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Schuller, B., Steidl, S., Batliner, A., Bergelson, E., Krajewski, J., Janott, C., Amatuni, A., Casillas, M., Seidl, A., Soderstrom, M., et al. (2017). The interspeech 2017 computational paralinguistics challenge: Addressee, cold & snoring. In Computational Paralinguistics Challenge (ComParE), Interspeech 2017.

Vos, T., Allen, C., Arora, M., Barber, R. M., Bhutta, Z. A., Brown, A., Carter, A., Casey, D. C., Charlson, F. J., Chen, A. Z., et al. (2016). Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990–2015: a systematic analysis for the global burden of disease study 2015. *The Lancet*, 388(10053):1545–1602.

Wu, J., Yu, Y., Huang, C., and Yu, K. (2015). Deep multiple instance learning for image classification and auto-annotation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3460–3469.

Zhang, C., Platt, J. C., and Viola, P. A. (2006). Multiple instance boosting for object detection. In Advances in neural information processing systems, pages 1417–1424.