Early Text Classification using Multi-Resolution Concept Representations

A. Pastor López-Monroy[⋆], **Fabio A. González**[†], **Manuel Montes-y-Gómez**[‡§],
Hugo Jair Escalante[‡] and **Thamar Solorio**[⋆]
[⋆] Department of Computer Science, University of Houston,Texas USA
[†] Systems and Computer Engineering Department, Universidad Nacional de Colombia
[‡] Instituto Nacional de Astrofísica, Óptica y Electrónica, Puebla México
[§] PRHLT Research Center, Universitat Politècnica de València, Spain
alopezmonroy@uh.edu, fagonzalezo@unal.edu.co,
{mmontesg, hugojair}@ccc.inaoep.mx, solorio@cs.uh.edu

Abstract

This paper proposes a novel document representation, called Multi-Resolution Representation (MulR), to improve the early detection of risks in social media sources. The goal is to effectively identify the potential risk using as little evidence as possible and with as much anticipation as possible. MulR allows us to generate multiple "views" of the text. These views capture different semantic meanings for words and documents at different levels of granularity, which is very useful in early scenarios to model the variable amounts of evidence. The experimental evaluation shows that MulR using low resolution is better suited for modeling short documents (very early stages), whereas large documents (medium/late stages) are better modeled with higher resolutions. We evaluate the proposed ideas in two different tasks where anticipation is critical: *sexual predator detection* and *depression detection*. The experimental evaluation for these early tasks revealed that the proposed approach outperforms previous methodologies by a considerable margin.

1 Introduction

Everyday there is a huge amount of people interacting in many social media sites. Unfortunately this immense cyber-world has been misused by cyber-criminals, who hide in the depths of the web. For this reason, the social media information has been increasingly studied in the context of applications related to security, forensics and e-commerce. Recently the early prediction scenarios have attracted the attention of the scientific community (Losada et al., 2017), which aims to prevent major threats in a number of practical situations by analyzing the text as evidence (e.g., sexual harassment, cyberbullying, etc).

In Natural Language Processing this emerging field is called early text classification and the goal is to identify risky-target categories by using as few text as possible and with as much anticipation as possible. In real scenarios the amount of evidence available from users under analysis is continuously growing. Consider for instance chat rooms, or posts and comments in social networks, these text sources comprise cumulative evidence for early prediction that can be used to better capture the phenomenon under study (Escalante et al., 2017; Losada et al., 2017). This scenario has challenging particularities. For example, in early stages where 10% or 20% of the information is available it is necessary to model very short length documents, which tend to produce sparse and low discriminative representations. On the other hand late stages require to exploit as much evidence as possible to make accurate predictions. This dynamism between the document length and classification stages makes necessary an adequate representation, that naturally copes with the dynamic amount of evidence in short and long texts generated by users at each stage. Traditional textual representations, such as Bag-of-Words (BoW) (Joachims, 1998), have problems dealing with social media short texts since they cause the representation to be high dimensional and very sparse. Moreover, in the particular case of early risk prediction, class unbalance and noisy text also represent a challenge.

In this paper we propose a representation that deals with these challenges by taking advantage of word vectors into a novel methodology for representing documents. This representation generates high-level features, that we called meta-words, which capture concepts at different resolution levels. A meta-word is a primitive construction represented by a vector that summarizes the information of semantically related words. Our methodology associates words with similar semantic meaning to the same meta-words. These meta-words

Proceedings of NAACL-HLT 2018, pages 1216–1225
New Orleans, Louisiana, June 1 - 6, 2018. ©2018 Association for Computational Linguistics

are obtained by applying clustering techniques to word representations, where the resultant "centroids" comprise the meta-words. Documents are then represented by a Bag-of-Centroids (BoC), that is, a histogram accounting for the occurrence of coarse thematic/semantic primitives, i.e., the meta-words. This part of the work is inspired by the Bag-of-Visual-Words (BoVW), which is widely used in computer vision to represent images (Sivic and Zisserman, 2004; Lazebnik et al., 2006).

The key aspect for early scenarios is that the number and size of meta-words, allow us to manipulate the level of granularity or the *resolution* of the representation. This property is very useful to capture discriminative information along the growing amount of available evidence at each early stage. We thus propose a multi-resolution approach, in which primitives at different resolutions are combined to capture feature concepts at multiple levels of detail. The contributions of this paper are twofold: (i) a new Multi-Resolution (MulR) document representation, a generalization to represent documents by exploiting word-vectors at different levels of resolutions; (ii) an empirical validation of the usefulness of multiple resolution levels for early risk detection on social media documents. Our experimental results show that this approach is a promising alternative for early text classification scenarios, where there is a need to make predictions as soon as possible, with little evidence, while at the same time, being robust to incorporate more evidence as it becomes available. We recorded experimental results of an extensive evaluation of our proposed techniques over two benchmarks for early scenarios: sexual predator detection and depression detection. Results showed that in all cases our methodology outperforms state-of-the-art methodologies.

Interestingly, document representations based on partitioning the word-embedding space, like ours, are somewhat similar to topic modeling based representations. In the experimental section we also compare the performance of our method to different topic-based representations like Latent Semantic Analysis (LSA) (Deerwester et al., 1990) and Latent Dirichlet Allocation (LDA) (Blei et al., 2003). Experimental results showed that our method outperformed the reference techniques. We elaborate on the benefits and limitations of our proposed techniques later in this paper.

2 Related Work

The Early Text Categorization problem is an emerging research topic with scant work (Dulac-Arnold et al., 2011; Escalante et al., 2016, 2017). Recently, the relevance of the problem has motivated specialized forums such as eRisk-CLEF17 (Losada et al., 2017). One of the first attempts is based on processing documents in a sentence-level basis (Dulac-Arnold et al., 2011). At every time t, the method reads a sentence and attempts to determine the class of the document. The key aspect of the work is a Markov Decision Process (MDP), where each sentence is modeled in a TFIDF vector. More recently, (Escalante et al., 2016) proposed a straightforward solution for early detection scenarios by using the naïve Bayes classifier. The idea consists in training with full documents, but when partial information has to be classified, the maximum a posteriori probability was estimated over the available text. Using this simple yet effective approach, the authors obtained competitive performance with the method in (Dulac-Arnold et al., 2011). Furthermore, results reported in (Escalante et al., 2016) were the first evaluation on early sexual predator detection.

In (Escalante et al., 2017) the authors propose methods to exploit Profile Based Representations (PBR's) for words (López-Monroy et al., 2015). PBRs are Distributional Term Representations of terms in the vocabulary. Similar to word embeddings these representations build a vector for each word, which aim to extract/learn concepts from simple occurrence statistics of terms in the target classes. PBRs capture discriminative information in a very low dimensional and non-sparse space suitable for early text classification problems. In other work, (Errecalde et al., 2017) successfully adapted a version of PBR's for the problem of early depression detection in the context of the eRisk-CLEF17 shared task. The evidence about PBRs suggests that this representation can naturally cope with missing information and obtain discriminative representations for incomplete documents. Nevertheless, just as the vast majority of word embeddings in the literature for standard text classification, there is no consensus about how to exploit these term vectors to represent entire phrases or documents (e.g., the most common strategy is to average the term vectors in documents).

The proposed method is based on creating meta-words to represent documents. Clustering words into meaningful groups based on some measure of similarity to represent text is not a new concept. One of the classic approaches is term clustering in an *unsupervised* manner that was first investigated by (Lewis, 1992). He called his method *reciprocal nearest neighbor clustering*. His method consists of joining words that are similar according to a measure of similarity. In other work, Brown et al. (1992) explored the idea of discovering similarities between words to obtain clusters at different levels. One key difference with our proposal is that in (Brown et al., 1992), terms are deterministically/probabilistically associated with a discrete class, where terms that are in the same class are similar in some aspect. However in our proposed strategy, we exploit word vectors instead of a discrete random deterministic variable (e.g., soft/hard partitions of word sets). This makes possible to discover different clusters and meta-words if we change the word representation. Thus, the proposed strategy is highly adaptable to other domains, where the specialization would be achieved by changing the word representation for the problem. In other work, Li and Jain (1998) found that term grouping helps to reduce the feature dimensionality, and at the same time, overcomes the generalization problem of feature selection. The evidence has showed that the performance of the classifier is, at least maintained (Li and Jain, 1998; Slonim and Tishby, 2001). Finally, other authors have also studied the problem of term clustering under a *supervised* scheme. For example, Baker and McCallum (1998) used a *supervised* scheme to cluster similar words. They carried out experiments using a Naive Bayes classifier and found results improvement by using a single word representation.

The methods proposed in this research work follow a line of thinking focused on the document representation rather than term representation. Hence, the proposed method takes advantage of specialized vector representation of words (e.g., PBR), but several extensions can be envisioned using other word embeddings in the literature. The benefits of our approach are that it is model independent, easy to implement, and computes lower dimensional and less-sparse representations than traditional BoW. More important, our method improves over state of the art methods, outperforming the methods in (Errecalde et al., 2017; Escalante et al., 2017) that in turn, outperform that in (Dulac-Arnold et al., 2011; Escalante et al., 2016).

3 Multi-Resolution Document Representation

We propose a multi resolution representation that allows to generate multiple "views" of the analyzed document. The intuition behind the proposal of a multi-resolution representation is that words will activate differently each view according to the amount of available text. We assume that having different resolution levels will allow to effectively represent the content of short and large texts as needed along different early stages. The proposed multi-resolution framework is depicted in Figure 1. The idea consists in associating words with similar meaning to the same meta-words in each resolution space. Documents are then represented by multiple Bag-of-Centroids (BoC), that is, multiple histograms accounting for the occurrence of coarse concepts. Hence, this representation can be seen as multiple BoW representations that incorporate multiple semantic resolutions. In Section 3.1 we describe the process to build a Bag-of-Centroids at a single resolution, then in Section 3.2 we formally present the Multi-Resolution variant.

3.1 Single Resolution: Bag of Centroids

Let $\mathcal{D} = \{(d_1, y_1), \ldots, (d_h, y_h)\}$ be a training set of h-pairs of documents d_i and class labels y_i. Also let $\mathcal{V} = \{w_1, \ldots, w_r\}$ denote the vocabulary of terms (in our case words). In order to create the Bag of Centroids (***BoC***) representation of each document, we first compute the vector representation v_i of each word w_i in the vocabulary of the collection. Note that our framework is agnostic to the underlying process for learning word representations and therefore any word vector representation can be used, for example word embeddings (Mikolov et al., 2013) or distributional term representations (Lavelli et al., 2004).

The proposed framework is based on the idea of clustering words using the semantic "distance" in the word embedding space. Thus, the first step of the algorithm consists of clustering the word embedding vectors v_i and finding the cluster centers to create the proposed meta-words. The representation of the vocabulary collection in the word embedding space W is the input for the clustering

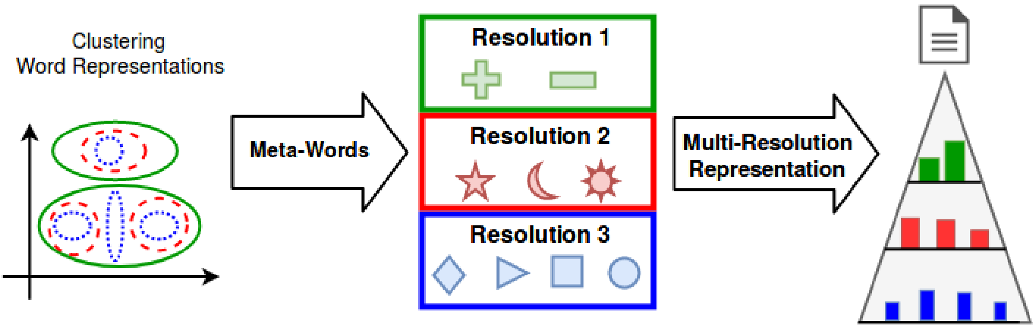

Figure 1: Algorithm to represent documents as meta-words using three hypothetical resolutions. The document is represented using the "meta-words" defined by the clustering of the vector representations of the vocabulary.

algorithm. For this purpose a variety of clustering approaches can be used. In our experimental evaluation, we explored different algorithms and found out that k-means offers a good trade-off between performance and speed. We applied k-means to the W representation to find the center of the clusters $C = \{c_1, c_2, \ldots, c_k\}$, with k being the number of selected centroids. Then, based on these cluster centers and using l_1-norm, we found a one-to-one association of each word to the *closest* cluster center in the word embedding space. In other words, for each word v_i, we can find an associated cluster center or meta-word c_u with $u \in \{1, 2, \ldots, k\}$. We denote this mapping by $c_u = closest(v_i, C)$, where $closest$ returns the centroid in C with the minimum distance to v_i. Finally, the **BoC**$_k$ representation for each document d_j corresponds to **BoC**$_k(d_j) = \{(c_\ell, n_\ell)\}_{\ell=1\ldots k}$ where c_ℓ corresponds to each of the k centroids and $n_\ell = |\{v_i | \forall v_i \in d_j, c_\ell = closest(v_i, C)\}|$. In other words, **BoC**$_k(d_j)$ corresponds to a histogram of centroid frequencies, where each pair (c_ℓ, n_ℓ) represents a centroid (meta-word) and its corresponding frequency in the document.

The BoC algorithm depends on one parameter: the number of clusters used to represent each document. This parameter is associated with the level of semantic coarseness used in the representation. In this regard, coarseness refers to the level of meta-word inclusivity: the more words associated with a single meta-word, the coarser the representation. Conversely, with fewer words, the representation becomes more granular. Note that this representation has well known parallels in the extreme cases. When each word becomes a centroid, the resulting representation is equivalent to the typical BoW representation, whereas a coarser representation, with only one meta word, will be equivalent to having the average meta-word of the entire collection.

3.2 Multi-Resolution BoC

The above proposed framework is particularly suitable for incorporating multi-resolution processing, given that the main parameter is related to the granularity or coarseness of the representation. As we will show in our analysis, this property is useful for early scenarios, since few/coarse meta-words allow to better encode documents with little text, whereas many/granular meta-words are useful when more text become available. We propose to exploit this multi-resolution version of the **BoC** representation. In this extension of the basic algorithm, we use a partition of the word embedding space at multiple levels and concatenate them into a new representation. Combining the different granularities into a single representation results in a more robust document model that can help to capture different amounts of text as needed. Intuitively, the coarser levels sufficiently classify documents in early stages, while the more granular levels exploit the additional evidence from longer documents on late stages. We present quantitative and qualitative experiments that support this claim in two datasets: Sexual Predator Detection and Depression Detection.

We call this variation of the BoC representation Multi-Resolution-BoC (**MulR**). Formally: **MulR**$(d_j) = \{$**BoC**$_{k_1}(d_j) \cup$ **BoC**$_{k_2}(d_j) \cup \ldots \cup$ **BoC**$_{k_n}(d_j)\}$, where $\{k_1, k_2, \ldots k_n\}$ correspond to a set of granular levels. Figure 1 shows the general framework, graphically depicting the process involved in transforming a document into a representation based on meta-words. The figure also includes the process of multi-resolution modification described above. In the figure, the meta-words depicted with 'blue' represent the more granular clusters, and those depicted with 'green' represent the less granular clusters. The multi-resolution BoC variation improves the performance by combining the information present at various levels of granularity. Moreover, when documents are closely related, more fine grained features allow to capture finer details and therefore produces better text classification results. This multi-resolution approach combines the advantages of both approaches to create an overall more effective classification method.

4 Data collections

For experiments we considered the two data sets described in Table 2. The tasks are Sexual Predator Detection (SPD) and Depression Detection, where clearly early detection is crucial. For the former we used the only publicly available data set for sexual predator detection (Inches and Crestani, 2012). This data set was released in the context of the sexual predator identification task at PAN-CLEF'12 and comprises a large number of chat conversations that include real sexual predators. Thus, the task approached is that of identifying those conversations that potentially include a sexual predator, as in (Villatoro-Tello et al., 2012; Escalante et al., 2013, 2016). For the depression detection task we use the dataset presented in (Losada et al., 2017). In this dataset, each instance has the post history for a user, and depressed users were self-identified as having been diagnosed with depression.

5 Evaluation Framework

For our experiments, we lower case the text in documents and use words and punctuation marks as terms[1]. The representation obtained for each document is then processed by a Support Vector Machine (SVM) with a linear kernel.

For the evaluation of the *earliness* performance, we report the performance of the different methods when using increasing amounts of textual evidence (chunk by chunk evaluation). This evaluation allows to quantify prediction performance when using partial information in documents, and it is a strategy that has been used to evaluate early classification (Escalante et al., 2016; Errecalde et al., 2017; Losada et al., 2017). For the evaluation of performance we used the $f_1 = \frac{2 \times precision \times recall}{precision + recall}$ measure. This decision was made in agreement with previous work that reports this metric for the positive class (Errecalde et al., 2017). Please note that, contrary to other measures, such as accuracy, f_1 measure accounts for the class imbalance problem when only the positive class is analyzed. This is desirable for the data sets we consider as they are highly unbalanced.

Word-vector representations: As previously mentioned, the proposed MulR representation generalizes word-vector representations and thus can extend any representation that models each term in the vocabulary using a vector. For this purpose a wide variety of word embeddings or distributional term representations could be used. Both of them exploit the distributional hypothesis to build word vectors, nonetheless they differ in the strategy to capture the relevant information. In this work we use the widely used word2vec, but also other representations that have been used in recent works for these collections. In Table 1 we describe each of the word vector representations considered for this work[2].

Baselines: The main baselines in this work are methods based on the idea of topic modeling for text classification. Topic-based representations group words into topics defined by a set of related words[3]. Given the strong relation to our method we compare our proposal against Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA). Furthermore, we also compare with Bag-of-Words using Term Frequency Inverse Document Frequency, since it is a traditional baseline in text categorization tasks.

6 Experimental Results

In this section we report the experimental results for the MulR representation and the selected approaches from the state-of-the-art. In all the experiments we trained the reference classifier (SVM) using full-length documents in the training dataset. In the testing phase, each approach uses all the available information in each of the ten chunks (each chunk increases the available text in 10%). More specifically, we generate document representations starting with the first chunk, and then incrementally adding one chunk at a time. The

[1] We used terms with frequency higher than 10 in the training datasets.

[2] For distributional representations we used the framework at https://github.com/lopez-monroy/FeatureSpaceTree

[3] We empirically set to 200 the size of the concept space.

Word Representation	Description
W2V (Mikolov et al., 2013)	Word2Vec uses the Skip-ngram model to find word representations that are useful to predict the surrounding words of a sentence or a document. The method is efficient for learning high-quality vector representations of words from large amounts of text data. We empirically set to 200 the vector dimension.
DOR (Lavelli et al., 2004)	Document Occurrence Representation (DOR) captures the semantics of a word by observing occurrence distribution over documents in the corpus. DOR represents each word v_i as a vector $\mathbf{t_i} = \langle t_{i,1}, \ldots, t_{i,\lvert\mathcal{D}\rvert} \rangle$, where $\lvert\mathcal{D}\rvert$ is the number of documents in the training collection, and $t_{i,k}$ indicates the relevance of the document D_k to characterize v_i.
TCOR(Lavelli et al., 2004)	In Term Co-occurrence Representation (TCOR) the semantics of a word is captured by observing its co-occurrences with other words across documents in the corpus. Thus, each word v_i is associated to a vector $\mathbf{t_i} = \langle t_{i,1}, \ldots, t_{i,\lvert\mathcal{V}\rvert} \rangle$, where $\lvert\mathcal{V}\rvert$ indicates the vocabulary size, and $t_{i,k}$ denotes the contribution of the word v_k to the semantic description of v_i.
PBR (López-Monroy et al., 2015)	Profile Based Representation exploits occurrence-statistics of words over a set of documents in target categories. PBR represents each word $v_i \in V$ with a vector $\mathbf{t_i} = \langle t_{i,1}, \ldots, t_{i,q} \rangle$, where the $t_{i,k}$ is the degree of association between word v_i and category C_k. The target categories can be taken from the task or artificially created by means of clustering such as in (Escalante et al., 2017).
TVT (Errecalde et al., 2017)	Temporal Variation Terms (TVT) is an adapted version of PBR for early scenarios. TVT builds new artificial target classes/labels in the training set simulating a text stream to generate enriched representations of $\mathbf{t_i}$. The idea is to exploit the positive category to create a set of new artificial categories using text-fragments.

Table 1: Word vector representations for early experimentation.

Task	Data set	Training	Test
Sexual predator det.	PAN'12	6588	15329
Depression det.	eRisk'17	486	401

Table 2: Data sets considered for early experimentation. There are only two classes in each dataset.

models will then make predictions incrementally as well. We report f_1 performance when using different amounts of text from test documents. For the proposed MulR representation, we build 5 different resolutions: 10, 50, 100, 500, and 1000. The goal was to generate meta-words at different levels of granularity, and we plan to further explore the impact of these resolutions in our future research.

In the following experiments, we used the word representations in Table 1 to build our proposed MulR document representation. For comparison purposes we also generate an alternative document representation by averaging (Avg) term-vectors of words in each document, which is a popular strategy to build document representations. Finally, we also compare against several traditional baselines such as the Bag-of-Words, LSA, and specialized methods in each collection (Escalante et al., 2017; Errecalde et al., 2017). We evaluate the usefulness of all these different representations in the two early classification tasks mentioned earlier.

6.1 Sexual Predators Detection

In this section we evaluate the performance of the proposed MulR and other reference methodologies for the SPD early detection task (Figure 2). We also show results for MulR and different word representations in Table 3, where several findings can be outlined. First of all, results obtained in early stages (chunk 1 to 4) using the proposed MulR are clearly superior to those obtained averaging word vectors. This is an interesting outcome, since the MulR representation seems to be useful for early scenarios independently of the word vector representation. In the particular case of MulR(TVT), the representation obtains an outstanding performance when having little information (e.g., performance between ≈71% and ≈90% before reading 50% of the text). More important, performance improves as more evidence is available (i.e., see the steady improvement up to ≈97%). These results show that MulR is a robust representation, even in the presence of different amounts of textual evidence, with a clear advantage for early classification stages.

In Figure 2 we can also observe that MulR representation outperformed, by a large margin, the proposed baselines; BoW-TFIDF, LSA, LDA. Furthermore, MulR representation obtains better performance than the work in (Escalante et al., 2017), which consists in averaging the PBRs (same that Avg-PBR) and is the state-of-the-art in early SPD. Note that different than (Escalante et al., 2017), the proposed MulR significantly improves even after reading 40% of the information. The experimental results in Table 3 also show the

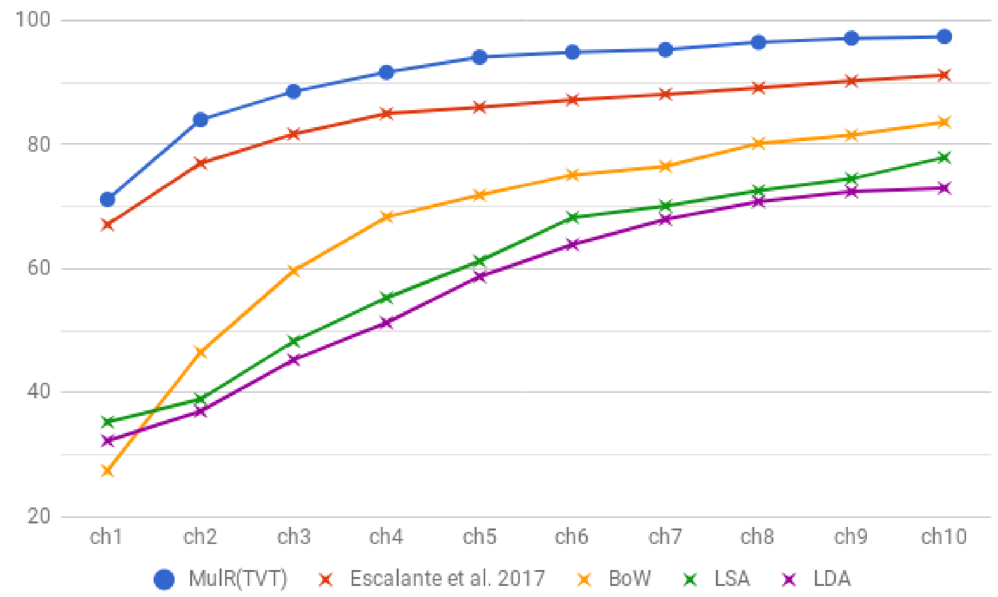

Figure 2: F_1 scores for the chunk by chunk evaluation of the reference methodologies in Sexual Predator Detection.

Method	ch_1	ch_2	ch_3	ch_4	ch_5	ch_6	ch_7	ch_8	ch_9	ch_{10}
BoW-TFIDF	27.40	46.51	59.62	68.33	71.84	75.11	76.49	80.19	81.51	83.63
LSA	35.22	38.93	48.25	55.27	61.21	68.24	70.12	72.54	74.49	77.91
LDA	32.22	36.98	45.27	51.27	58.70	63.87	67.94	70.81	72.41	72.98
Avg(W2V)	55.74	63.87	70.11	82.53	87.24	88.97	88.01	87.45	84.71	83.12
MulR(W2V)	58.97	65.78	71.97	83.09	85.49	87.21	88.46	89.00	89.15	89.49
Avg(DOR)	66.71	76.54	81.01	91.14	92.23	93.91	95.19	95.87	96.47	96.59
MulR(DOR)	68.24	78.77	87.14	**92.07**	**94.18**	94.04	94.84	95.24	95.46	95.97
Avg(TCOR)	60.17	67.97	74.41	78.51	81.24	82.71	83.97	82.51	82.90	82.27
MulR(TCOR)	61.51	69.12	75.43	78.89	80.26	79.97	81.01	81.59	82.14	83.01
Avg(PBR)	67.10	76.97	81.69	85.00	86.03	87.21	88.14	89.16	90.25	91.21
MulR(PBR)	69.16	77.41	83.01	87.05	88.07	89.27	90.14	91.51	92.01	92.41
Avg(TVT)	65.74	80.24	86.19	90.25	92.02	93.13	94.39	95.23	95.89	96.58
MulR(TVT)	**71.15**	**84.00**	**88.56**	91.66	94.11	**94.92**	**95.31**	**96.50**	**97.16**	**97.43**
(Escalante et al., 2017)	67.10	76.97	81.69	85.00	86.03	87.21	88.14	89.16	90.25	91.21

Table 3: F_1 results for the chunk by chunk evaluation of different approaches in Sexual Predator Detection. The proposed MulR is evaluated using different word vector representations in the literature.

following interesting findings:

1. The most useful word vector representation is TVT (Errecalde et al., 2017). This is not surprising, since TVT is a specialized distribution term representation for early prediction scenarios.

2. Word2Vec [4] representations obtained moderate performance in all experiments. We infer that much more data of these specific social media domains are needed in order to build suitable models.

3. MulR representation is an effective solution for all early chunks, but as more text is available, the other methodologies significantly increase their discriminative power, as seen in results for later chunks. In fact, some representations such as Avg(DOR) can outperform MulR(DOR) representation in late stages. However, even under these conditions MulR(TVT) and MulR(DOR) outperform all reference methodologies.

6.2 Depression Detection

In Table 4 we show the experimental results for early depression detection. In Figure 3 we highlight the performance of the proposal and the reference methodologies. From these results we point

[4] Embeddings were trained in each dataset. We tested pre-trained word embeddings for wikipedia/twitter, but the performance was worse.

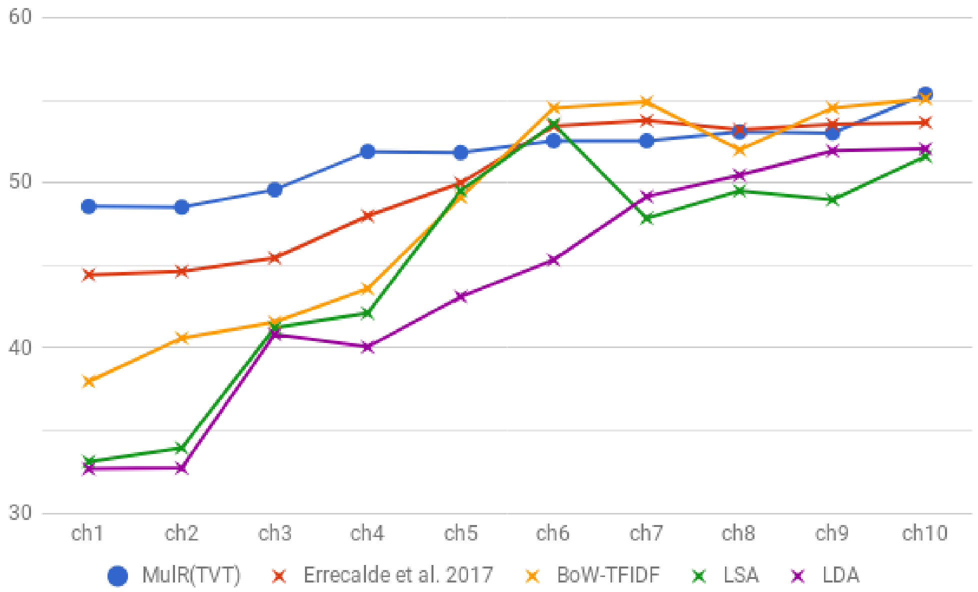

Figure 3: F_1 scores for the chunk by chunk evaluation of the reference methodologies in Depression Detection.

Method	ch_1	ch_2	ch_3	ch_4	ch_5	ch_6	ch_7	ch_8	ch_9	ch_{10}
BoW-TFIDF	37.97	40.60	41.56	43.59	49.12	54.55	54.90	52.00	54.55	55.10
LSA	33.12	33.94	41.22	42.11	49.52	53.57	47.86	49.52	48.98	51.61
LDA	32.68	32.73	40.83	40.06	43.11	45.33	49.17	50.47	51.94	52.06
Avg(W2V)	35.86	41.03	47.06	48.25	44.93	51.80	54.38	56.14	55.28	55.14
MulR(W2V)	41.34	43.79	47.20	48.44	47.67	52.00	53.91	54.18	54.29	54.39
Avg(DOR)	46.06	47.23	48.02	50.54	54.26	58.27	**57.81**	**58.73**	**59.84**	**66.12**
MulR(DOR)	47.55	48.12	48.38	51.83	**55.56**	**58.49**	52.43	53.06	57.73	54.35
Avg(TCOR)	37.42	44.44	44.60	48.64	49.64	53.33	52.94	53.44	52.46	58.32
MulR(TCOR)	44.76	47.95	46.81	48.32	51.47	52.11	54.01	54.55	56.30	57.06
Avg(PBR)	36.70	45.71	44.00	47.83	46.67	51.61	51.69	52.27	49.41	51.76
MulR(PBR)	40.98	46.15	44.83	48.70	50.00	53.10	57.39	54.55	54.55	55.86
Avg(TVT)	39.18	44.21	45.83	46.94	48.42	51.02	48.94	46.15	48.35	51.11
MulR(TVT)	**48.57**	**48.53**	**49.59**	**51.90**	51.83	52.55	52.55	53.09	53.03	55.38
(Errecalde et al., 2017)	44.44	44.64	45.45	48.00	50.00	53.44	53.77	53.23	53.55	53.66

Table 4: F_1 results for the chunk by chunk evaluation of different approaches in Depression Detection. The proposed MulR is evaluated using different word vector representations in the literature.

out several interesting findings. The first one is that for this collection, results obtained by the proposal are clearly superior to others in early stages. In general, we can observe the following:

1. The most useful representation in early stages was MulR(TVT), which have considerable improvements between ≈5% and ≈2% in chunks 1 to 4.

2. Word Embeddings and DOR showed a similar behavior than in SPD. But in late stages, the best representation was Avg(DOR).

3. Depression Detection problem is a much harder problem than SPD. The F_1 measure is under ≈60% in most of the results. This could be due to the highly unbalanced dataset in two ways: i) the number of instances in each class, and ii) the amount of text contained in documents.

6.3 The Relevance of Individual Resolutions

In this section we aim to study the role of the different resolutions in early scenarios.[5] The purpose of the first analysis is to observe the performance of each individual resolution in MulR. In Table 5 we show the results of MulR(TVT) under each of the five resolutions $(R_1 = 10, R_2 = 50, R_3 = 100, R_4 = 500, R_5 = 1000)$ and each chunk.

[5] The number and size of resolutions, could improve the performance, but it is a future research path to enhance the characterization of specific data sets.

Sexual Predator Detection										
Method	ch_1	ch_2	ch_3	ch_4	ch_5	ch_6	ch_7	ch_8	ch_9	ch_{10}
MulR(TVT)-R1	59.88	74.71	82.17	85.82	89.19	90.19	91.51	92.61	92.42	92.51
MulR(TVT)-R2	70.03	83.03	88.04	90.49	93.00	94.01	94.85	94.70	94.90	95.02
MulR(TVT)-R3	67.87	82.41	87.23	90.17	92.07	92.88	93.91	94.89	95.49	96.20
MulR(TVT)-R4	66.10	80.33	85.89	88.76	90.48	91.86	93.14	93.85	94.73	95.38
MulR(TVT)-R5	62.34	77.73	83.05	86.72	88.64	90.32	92.16	93.04	93.00	93.76
MulR(TVT)	**71.15**	**84.00**	**88.56**	**91.66**	**94.11**	**94.92**	**95.31**	**96.50**	**97.16**	**97.43**

Table 5: F_1 results for the chunk by chunk evaluation of different approaches in Sexual Predator Detection. The best MulR(TVT) is separately evaluated under each resolution.

For early SPD the evidence is clear; as the resolution increases the performance in early stages decrease.[6] Also note that the higher the resolution, the more chunks needed to outperform the result of the previous resolution. For example, resolution R_3 outperforms R_2 in chunk 8. Also note that R_4 and R_5 needed more chunks to obtain comparable performance than R_2. Our experimental results excluding one resolution at the time showed worse performance, therefore all of them are essential in the overall classification. Clearly, this evidence shows that the MulR representation is in fact very useful.

Sexual Predator Detection					
Test set	R_1	R_2	R_3	R_4	R_5
chunk-1	4	3	2	1	0
chunk-2	3	2	3	2	0
chunk-3	3	2	2	2	1
chunk-4	3	2	3	1	1
chunk-5	3	2	3	1	1
chunk-6	3	1	4	1	1
chunk-7	3	1	3	2	1
chunk-8	2	2	1	2	3
chunk-9	3	1	1	2	3
chunk-10	3	0	2	2	3

Table 6: Post-analysis in test dataset. Distribution of the top ten meta-words according to each resolution R_i at different chunks. We used Information Gain (Hall et al., 2009) to rank meta-words in MulR(TVT).

In Table 6 we provide further evidence about the role of different resolutions. In this complementary analysis we study each chunk at test data. For this we use the MulR learned in training to represent test documents, then we compute the Information Gain using Weka (Hall et al., 2009) at each test chunk. In Table 6 we show the number of features in each resolution R_i that are present in the top ten meta-words of the MulR(TVT). The analysis complements the evidence, lower resolutions have higher IG at early chunks, whereas higher resolutions are more necessary in late chunks.

7 Conclusions

In this paper we proposed a multi resolution representation that allows to generate multiple "views" of the document. Intuitively these views expose different semantic meanings for words and documents along different resolutions. The different resolutions allow to effectively represent the content of short and large texts at different early stages. The MulR obtained the best results reported so far on the early Sexual Predator Detection task dataset (Inches and Crestani, 2012). For Depression Detection the chunk by chunk evaluation shows promising results for MulR in early stages. What is more, it was shown that the MulR further improves the early recognition performance in the two tasks using different word representations. The relevance of the resolutions in these results is a key factor to understand the proposed MulR and future extensions. These results provide solid evidence to further research on this topic and encourage researchers to apply and evaluate the usefulness of multi-resolution features for other related *early* tasks.

Acknowledgments

This research was partially supported by CONACYT-México (project FC-2016/2410).

References

Douglas Baker and Andrew Kachites McCallum. 1998. Distributional Clustering of Words for Text Classification. In *Proceedings of the 21st ACM International Conference on Research and Development in Information Retrieval*. pages 96–103.

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

Peter F. Brown, Peter V. deSouza, Robert L. Mercer, Vincent J. Della Pietra, and Jenifer C. Lai.

[6] The only exception to this is R_1, which has the lowest overall performance. This is somewhat expected since this space only has 10 features to represent documents.

1992. Class-based n-gram models of natural language. *Comput. Linguist.* 18(4):467–479.

S. Deerwester, S. Dumais, T. Landauer, G. Furnas, and R. Harshman. 1990. Indexing by Latent Semantic Analysis. *Journal of the American Society of Information Science* 41(6):391–407.

G. Dulac-Arnold, L. Denoyer, and P. Gallinari. 2011. Text classification: A sequential reading approach. In *Advances in Information Retrieval, Proc. of 33rd European Conference on IR Research, (ECIR'11)*. Springers, volume 6611 of *LNCS*, pages 411–423.

Marcelo L. Errecalde, Ma. Paula Villegas, Dario G. Funez, Ma. José Garciarena Ucelay, and Leticia C. Cagnina. 2017. Temporal variation of terms as concept space for early risk prediction. In *CLEF (Working Notes)*.

H. J. Escalante, A. Juarez, E. Villatoro, M. Montes-y-Gómez, and L. Villasenor. 2013. Sexual predator detection in chats with chained classifiers. In *Proceedings of NAACL-HLT 2013, 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. page 46.

H. J. Escalante, M. Montes-y-Gómez, L. Villasenor, and M. L. Errecalde. 2016. Early text classification: a naive solution. In *Proceedings of NAACL-HLT 2016, 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. pages 91–99.

H. J. Escalante, E. Villatoro-Tello, S. E. Garza, A. P. López-Monroy, M. Montes-y-Gómez, and L. Villaseñor-Pineda. 2017. Early detection of deception and aggressiveness using profile-based representations. *Expert Systems with Applications* 89(Supplement C):99 – 111.

M. Hall, E. Frank, G. Holmes, B. Pfahringer, P Reutemann, and I. Witten. 2009. The weka data mining software: An update. *SIGKDD Explorations* 11.

G. Inches and F. Crestani. 2012. Overview of the international sexual predator identification competition at pan-2012. In *CEUR Workshop Proceedings, Working Notes for CLEF 2012 Conference*. CEUR, volume 1178.

Thorsten Joachims. 1998. Text categorization with support vector machines: Learning with many relevant features. In Claire Nédellec and Céline Rouveirol, editors, *Machine Learning: ECML-98*, Springer Berlin Heidelberg, volume 1398 of *Lecture Notes in Computer Science*, pages 137–142.

Alberto Lavelli, Fabrizio Sebastiani, and Roberto Zanoli. 2004. Distributional term representations: an experimental comparison. In *CIKM*. ACM, pages 615–624.

Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. 2006. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. volume 2, pages 2169–2178.

David D. Lewis. 1992. An Evaluation of Phrasal and Clustered Representations on a Text Categorization Task. In *Proceedings of the 15th International ACM/SIGIR Conference on Research & Development in Information Retrieval*. June, pages 37–50.

Yong Li and Anil Jain. 1998. Classification of Text Documents. *The Computer Journal* 41(8):537 –546.

A. Pastor López-Monroy, Manuel Montes-y-Gómez, Hugo Jair Escalante, Luis Villasenor-Pineda, and Efstathios Stamatatos. 2015. Discriminative subprofile-specific representations for author profiling in social media. *Knowledge-based Systems* 89:134–147.

David E. Losada, Fabio Crestani, and Javier Parapar. 2017. *eRISK 2017: CLEF Lab on Early Risk Prediction on the Internet: Experimental Foundations*, Springer International Publishing, Cham, pages 346–360.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Proceedings of Advances in Neural Information Processing Systems 26 (NIPS 2013)*. pages 1–9.

J. Sivic and A. Zisserman. 2004. Video data mining using configurations of viewpoint invariant regions. In *CVPR*. IEEE, volume 1, pages I–488.

Noam Slonim and Naftali Tishby. 2001. The Power of Word Clusters for Text Classification. In *Proceedings of the 23rd European Colloquium on Information Retrieval Research*. volume 1, pages 1–12.

Esaú Villatoro-Tello, Antonio Juárez-González, Hugo Jair Escalante, Manuel Montes-y Gómez, and Luis Villaseñor Pineda. 2012. A two-step approach for effective detection of misbehaving users in chats. In *CLEF (Online Working Notes/Labs/Workshop)*.