# UNSUPERVISED DETECTION OF VIOLENT CONTENT IN ARABIC SOCIAL MEDIA

Kareem E Abdelfatah[1,3], Gabriel Terejanu[1],Ayman A Alhelbawy[2,3]

[1]Department of Computer Science and Engineering,
University of South Carolina, Columbia, SC, USA
[2]Computer Science and Electrical Engineering Department,
University of Essex, Essex, United Kingdom
[3]Computers and Information Faculty, Fayoum University, Fayoum, Egypt

### ABSTRACT

*A monitoring system is proposed to detect violent content in Arabic social media. This is a new and challenging task due to the presence of various Arabic dialects in the social media and the non-violent context where violent words might be used. We proposed to use a probabilistic non-linear dimensionality reduction technique called sparse Gaussian process latent variable model (SGPLVM) followed by k-means to separate violent from non-violent content. This framework does not require any labelled corpora for training. We show that violent and non-violent Arabic tweets are not separable using k-means in the original high dimensional space, however better results are achieved by clustering in low dimensional latent space of SGPLVM.*

### KEYWORDS

*Violence, Social Media, Arabic, SGPLVM, Dimensionality Reduction, Unsupervised learning*

## 1. INTRODUCTION

According to the Arab Social Media Report, there were 6 million Twitter users in the Arab world in March 2014, posting on average around 17 million tweets per day [1]. Twitter provides profound information as people share with others what they like and do not like, their beliefs, their political opinions, and what they observe. Due to dramatic problems plaguing much of the Arab world, a significant amount of content on social media is about violence and abuse.

Detecting offensive and violent content in social media is a very active research area, especially in the last few years. This type of research is valuable to various organizations such as Human Rights Organizations (HRO). In some crisis countries like Iraq or Syria, it may be dangerous and not safe for HROs to obtain reports and monitor the human rights situations through the usual process. Therefore, mining social media might be a solution to the problem of detecting and identifying human rights abuses safely. However, according to our knowledge there is very little work for detecting violent content in Arabic social media. This is a serious gap, as there is a real need for such kind of research in Arabic social media.

Arabic language in social media is one of the most challenges languages to be study and analyzed. Arabic is the official language in around 22 countries with more than 350 million people around the world [2]. All of these countries are Diglossia societies where both the standard form of the language, Modern Standard Arabic (MSA), and the regional dialects (DA) are used [3]. MSA is used in official settings while DA is the native tongue of Arabic speakers. DA does

Dhinaharan Nagamalai et al. (Eds) : CoSIT, SIGL, AIAPP, CYBI, CRIS, SEC, DMA - 2017
pp. 01– 07, 2017. © CS & IT-CSCP 2017 DOI : 10.5121/csit.2017.70401

not have a standard orthography and it is divided into several groups among these countries [4]. Nowadays, these dialects are extensively utilized in social media text, in spite of their original absence from a written form [3].

Detecting violence content in Arabic social media is not a trivial task. Not only because the different Arabic dialects that we have mentioned above, but also because of violent Arabic words are not always representative of violent context. For example, the word “Killing” has both a violent meaning but it may also be used in a non-violent context as in the following tweet examples [5].

إن الذاكرة والألم توأمان لا تستطيع قتل الألم  دون سحق الذاكرة

"The memory and the pain twins, you cannot kill the pain without crushing the memory"

تستطيع قتل الأزهار ولكن لا تستطيع أن تمنع قدوم الربيع

"You may kill the flowers but cannot prevent the arrival of spring"

On other hand, the same word can be used in a violent context, like the following example [5]:

مقتل خمسة أشخاص برصاص مسلحين والقبض علي ستة مشتبه بهم

"The killing of five people shot dead by gunmen and arrested six suspects"

In this work, we tackle this problem using a recently released dataset that contains 16234 manually annotated Arabic tweets [5]. It contains different violent context like killing, raping, kidnapping, terrorism, invasion, explosion, or execution, etc. According to our knowledge this is the first study conducted on this dataset. We use an unsupervised technique to binary cluster this dataset to violent and non-violent content. First, the Sparse Gaussian Process Latent Variable Model (SG- PLVM) [6] is used as an unsupervised probabilistic non-linear Dimensionality Reduction (DR) model. Then we apply k-means on the features extracted in the previous step. Using recent released Arabic dataset [5], our experiments show that violent and non-violent Arabic tweets are not separable using k-means in the original high dimensional space, however better results are achieved using low dimensional projections provided by the SGPLVM.

## 2. PREVIOUS WORK

There is much research work in detecting violent content on web [7, 8]. Computer vision techniques have been proposed to detect violence in videos [9–11]. On the other hand, text mining techniques have been used to detect violence in English social media; but little work targets this problem in Arabic social media.

A probabilistic violence detection model (VDM) is proposed in Ref. [12] to extract violence related topics from social media data. The authors propose a weakly supervised technique and they used OpenCalais with Wikipedia documents, and Wikipedia and YAGO categories to build a training corpus. The dataset was built to detect violence categories such as Crimes, Accidents, War Conflict, etc. Non-violence related categories are anything other than violence, like Education and Sports. We tested OpenCalais, but unfortunately it does not support Arabic text. Also, the number of documents under violence categories in Arabic Wikipedia is very small.

Lexical Syntactical Feature (LSF) [13] has been introduced to detect offensive language in social media. The proposed system uses the user profile to get some information about the user’s

English writing style. A set of lexical features like Bag of Words and N-grams, and hand-authoring syntactic rules are used to identify name-calling harassments. In additions, a users potentiality to send out offensive content in social media has been predicted using some features like style, structure, and context-specific features. This proposed method uses Naive Bayes and SVM techniques to train a classifier.

## 3. CLUSTERING IN A LOWER SPACE

It is very common in NLP to have a really high dimensional feature vectors. Using unsupervised techniques for clustering patterns is good and cheap choice. k-means algorithm is one of the good candidates for unsupervised learning techniques. But, k-means can give better results when it is applied on low dimensional features [14] Therefore, it is common to project a high dimensional data set onto a lower dimensional subspace using unsupervised DR techniques such as Principle Components Analysis (PCA) [15] to improve learning. It is widely used approach to project data onto a lower dimensional subspace using PCA then use k-means to cluster the data in the lower dimensions space [15].

Because unsupervised clustering algorithms such as k-means operate mainly on distances, it is vital to use a DR technique that is able to preserve the distance metric between the data points in the low dimensional subspace. PCA is the most widely used linear DR for obtaining a lower dimensional representation of a data set. PCA may maintain the dissimilarity [14] which can help the K-means to achieve better separation for clustering. We meant by preserve the dissimilarity is the ability to preserve the points that are far apart in data space to be far apart in the latent space. However, due to linearity, PCA may not capture the structure of the data through a low dimensional embedding [16].

Gaussian process latent variable model (GPLVM) [17] is a flexible non-linear approach to probabilistic modelling data in high dimensional spaces. It can be used as DR method which maps between the observed data points $Y \in \Re^{N\times D}$ and latent unobserved data points $X \in \Re^{N\times q}$ . One of its advantages it can preserve the dissimilarity and smoothness between the data in high and low dimension spaces. Smoothness means that if two points in the latent space are close (far) to each other then they will be mapped to two points that are relatively close (far) to each other in the data space. The GPLVM as a probabilistic approach models the relationship between latent variables and the observed data through non-linear parametrized function $y_{:,\mathrm{i}} = \mathrm{f}(\mathrm{X}, \mathrm{w}_{\mathrm{i},:}) + \varepsilon_{:,i}$ where $y_{:,\mathrm{i}} \in \Re^{N\times 1}$ represents one dimension of the observed data and $w_{\mathrm{i},:} \in \Re^{1\times D}$ is one row of the parameters $W \in \Re^{q\times D}$ which it has a prior Gaussian distribution over each of its row with zero mean and unit variance $w_i \sim N(w_i \mid 0, I)$and noise $\varepsilon_{:,\mathrm{i}} \sim N(0, \sigma^2 I)$. GPLVM assumes that there is independency across the data dimensionality. Thus, the likelihood for all dimensions can be written as a product of the likelihood of the $D$ observed dimensions.

$$p(Y \mid X) = \prod_{\mathrm{i}=1}^{D} N\left(y_{:,\mathrm{i}} \mid 0, \mathrm{K} + \sigma^2 I\right)$$

Inferencing the latent projects can be achieved through maximizing the marginal log-likelihood of the data,

$$\log p(Y \mid X) = \frac{-D}{2}\log|K| - \frac{1}{2}\mathrm{Tr}\left(K^{-1}\mathrm{YY}^{T}\right) + \mathrm{C}$$

Here, C is a constant and $K \in \Re^{N \times N}$ is a kernel matrix that is calculated from the training data. There are different kernel functions available that can be used. In our experiments we used the radial basis function (RBF),

$$k\left(x_i, \mathrm{x}_j\right) = \theta_{\mathrm{rbf}}\exp\left(\frac{-\left(x_i - x_j\right)^{T}\left(x_i - x_j\right)}{2\gamma^2}\right)$$

where $\theta_{rbf}$, $\gamma$ are the parameters or the kernel.

However, a major drawback with the standard GPLVM approach is its computational time. To infer the latent variable *X*, GPLVM uses a gradient based iterative optimization of the log likelihood which requires $O\left(N^3\right)$ complexity due to the inverse of K [6]. Therefore, the Sparse-GPLVM (SGPLVM) [6] comes to solve this issue by reducing the complexity to $O(u^2N)$ where u is the number of points retained in the sparse representation. Therefore, using Sparse-GPLVM before K-means can guarantee to preserve the dissimilarity between the data points in the latent space which leads to coherent patterns that can be detected easily via clustering.

## 4. DATASET

A manually annotated dataset of 16,234 tweets are used for training and testing [5]. Every tweet had been classified by at least five different annotators. As every tweet is classified by different users, it may be assigned different classes. So, a final aggregate class is assigned based on a class confidence score as it is described in the original publication [5]. In our experiments we have kept only the tweets have a confidence score more than 0.7.

*Table 1: Dataset Details*

| Class | Training | Testing | Total | % |
|---|---|---|---|---|
| Violence | 5673 | 2759 | 9332 | 57.5 |
| Non-Violence | 4790 | 2112 | 6902 | 42.5 |
| Total | 11363 | 4871 | 16234 | |

The original dataset is classified into seven violence classes: crime, violence, human rights abuse, political opinion, crisis, accidents, and conflict. There is an additional class “other”, which contains non-violence tweets where some violence words had been mentioned.

Because we are interested in detecting the violence acts in Arabic social media regardless the type of violence, all violence classes are mapped to one class “violence”, while the “other” class is mapped to “non-Violence” class. Around 70% of the dataset is used for training and 30% is used for testing as shown in Table 1.

## 5. EXPERIMENTS SETUP

The Arabic has a complex morphological structure especially for Dialectal Arabic [18]. Until now, there are no available standard resources for Arabic text mining and morphological analysis [18]. However for our study, we use MADIMARA [18] analysis tool because it has most of

common tools for morphological analysis in Arabic. After removing Arabic stop words and web links, we used MADIMARA to extract some morphological features like gloss and token.
Tweets are represented in a Vector Space Model (VSM) with TF-IDF weighting scheme The baseline approach is to cluster the dataset in the original high dimensional space into two clusters using k means [19] with different features. Then, two different DR techniques (PCA and SGPLVM) are applied. We study the ability to separate these data points in the latent space by clustering the data into two clusters using k-means. We have tried to reduce the data to different dimension (Dim) spaces and reported some of these results. Two sets of experiments have been carried out. The first set is using the Gloss feature where the original space is 14,621 dimensions. So we reduced it to 11,000, and 8000 with PCA.

Another experiment has been carried out with reducing dimensionality to 8000 but using GPLVM. The second set has been carried out using the token feature where the dimensionality is much higher i.e. 44,163 features. PCA had been used to reduce it to 35,000 and 8,000 features. For comparability reasons, GPLVM also used to reduce the dimensionality to 8000 again. To measure the performance for the clustering steps whether in the data space or latent space, we used the training data to assign each cluster to one class which maximizes the precision of “violence” class. Then, we use the precision (P), recall (R), and F-score (F) as an evaluation metric to compare between these different techniques.

## 6. RESULTS

Table 2 shows the results for applying k-means on the original data space and after reducing the dimensionality using PCA and SGPLVM with different features.

Gloss as a linguistic feature has a level of abstraction which multiple tokens may have the same gloss. It is noiseless as comparable to token feature each new token is considered as a new dimension.

From the gloss results, when we try to reduce the dimension using PCA to different levels (reduced to more than 45% of the original dimension), k-means is still able to sustain its performance. In two cases (higher and lower dimension), k-mean can achieve precision around 47%. However, in the token case, we can see that PCA is affected by the data representation which is noisy.

*Table 2: Experimental results.*

| **Gloss Feature** | | | | |
|---|---|---|---|---|
| Model | Dim | P | R | F |
| K-means (in data space) | 14,621 | 0.46 | 0.65 | 0.54 |
| PCA + K-means | 11,000 | 0.47 | 0.66 | 0.55 |
| PCA + K-means | 8000 | 0.46 | 0.63 | 0.54 |
| SGPLVMx + K-means | 8000 | **0.56** | 0.60 | **0.58** |
| **Token Feature** | | | | |
| Model | Dim | P | R | F |
| K-means (in data space) | 44,163 | 0.50 | 0.75 | 0.60 |
| PCA + K-means | 35,000 | 0.56 | 0.98 | 0.71 |
| PCA + K-means | 8000 | 0.49 | 0.72 | 0.58 |
| SGPLVMx + K-means | 8000 | **0.58** | 0.55 | 0.56 |

On the other hand, according to the precision metric, SGPLVM can help k-means to achieve better results than both of what it can obtain in the original space and in the lower space using PCA. Using SGPLVM for gloss features, we can increase the precision accuracy around 21% comparable to what we can get from other methods. However using token features, we tremendously decreased the dimensionality using SGPLVM to study if it is able to keep the distance between the points. Unlike the PCA, the results show that the non-linearity of SGPLVM with k-means is still able to outperform the k-means in the high data space.

## 7. CONCLUSION

In this paper we tackled a new challenging problem in Arabic social media. We introduced an unsupervised framework for detecting violence in the Arabic Twitter. We use a probabilistic non-linear DR technique and an unsupervised cluster algorithm to identify violent tweets in Arabic social media dataset. We compare k-means as a baseline with the results of SGPLVM and PCA with k-means. The preliminary results show that detecting violent content in this dataset using unsupervised techniques can be achieved using a lower dimensional representation of the data with results better than applying clustering on the original data set. More experiments will be carried out to achieve better results.

## REFERENCES

[1] R. Mourtada and F. Salem, "Citizen engagement and public services in the arab world: The potential of social media," Arab Social Media Report series,, 2014.

[2] F. Sadat, F. Kazemi, and A. Farzindar, "Automatic identification of arabic language varieties and dialects in social media," Proceedings of SocialNLP, 2014.

[3] H. Elfardy and M. T. Diab, "Sentence level dialect identification in arabic.," in ACL (2), pp. 456–461, 2013.

[4] N. Y. Habash, "Introduction to arabic natural language processing," Synthesis Lectures on Human Language Technologies, vol. 3, no. 1, pp. 1–187, 2010.

[5] A. Alhelbawy, P. Massimo, and U. Kruschwitz, "Towards a corpus of violence acts in arabic social media," in Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), (Paris, France), European Language Resources Association (ELRA), 2016.

[6] N. D. Lawrence, "Learning for larger datasets with the gaussian process latent variable model," in International Conference on Artificial Intelligence and Statistics, pp. 243–250, 2007.

[7] E. Whittaker and R. M. Kowalski, "Cyberbullying via social media," Journal of School Vio- lence, vol. 14, no. 1, pp. 11–29, 2015.

[8] A. Kontostathis, L. Edwards, and A. Leatherman, "Text mining and cybercrime, "TextMining: Applications and Theory. John Wiley & Sons, Ltd, Chichester, UK, 2010.

[9] E. B. Nievas, O. D. Suarez, G. B. Garc ́ıa, and R. Sukthankar, "Violence detection in video using computer vision techniques," in Computer Analysis of Images and Patterns, pp. 332–339, Springer, 2011.

[10] F. D. de Souza, G. C. Cha ́vez, E. A. do Valle, and A. de A Araujo, "Violence detection in video using spatiotemporal features," in Graphics, Patterns and Images (SIBGRAPI), 2010 23rd SIBGRAPI Conference on, pp. 224–230, IEEE, 2010.

[11] A. Datta, M. Shah, and N. D. V. Lobo, “Person-on-person violence detection in video data,” in Pattern Recognition, 2002. Proceedings. 16th International Conference on, vol. 1, pp. 433–438, IEEE, 2002.

[12] A. E. Cano Basave, Y. He, K. Liu, and J. Zhao, “A weakly supervised bayesian model for violence detection in social media,” in Proceedings of International Joint Conference on Natural Language Processing, pp. 109–117, Asian Federation of Natural Language Processing, 2013.

[13] Y. Chen, Y. Zhou, S. Zhu, and H. Xu, “Detecting offensive language in social media to protect adolescent online safety,” in Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Conference on Social Computing (SocialCom), pp. 71–80, IEEE, 2012.

[14] C. Ding and X. He, “K-means clustering via principal component analysis,” in Proceedings of the twenty-first international conference on Machine learning, p. 29, ACM, 2004.

[15] H. Zha, X. He, C. Ding, M. Gu, and H. D. Simon, “Spectral relaxation for k-means clustering,” in Advances in neural information processing systems, pp. 1057–1064, 2001.

[16] N. Lawrence, “Probabilistic non-linear principal component analysis with gaussian process latent variable models,” The Journal of Machine Learning Research, vol. 6, pp. 1783–1816, 2005.

[17] N. D. Lawrence, “Gaussian process latent variable models for visualisation of high dimensional data,” Advances in neural information processing systems, vol. 16, no. 3, pp. 329–336, 2004.

[18] N. Habash, R. Roth, O. Rambow, R. Eskander, and N. Tomeh, “Morphological analysis and disambiguation for dialectal arabic.,” in HLT-NAACL, pp. 426–432, 2013.

[19] J. A. Hartiganand M. A.Wong, “Algorithmas136:A k-means clustering algorithm,” Journal of the Royal Statistical Society. Series C (Applied Statistics), vol. 28, no. 1, pp. 100–108, 1979.