Open Research Online

The Open University's repository of research publications
and other research outputs

A weakly supervised Bayesian model for violence detection in social media

Conference or Workshop Item

How to cite:

Cano Basave, Amparo Elizabeth; He, Yulan; Liu, Kang and Zhao, Jun (2013). A weakly supervised Bayesian model for violence detection in social media. In: Sixth International Joint Conference on Natural Language Processing: Proceedings of the Main Conference, Asian Federation of Natural Language Processing, pp. 109–117.

For guidance on citations see FAQs.

© 2013 Asian Federation of Natural Language Processing

Version: Version of Record

Link(s) to article on publisher's website:
http://lang.cs.tut.ac.jp/ijcnlp2013/

Copyright and Moral Rights for the articles on this site are retained by the individual authors and/or other copyright owners. For more information on Open Research Online's data policy on reuse of materials please consult the policies page.

oro.open.ac.uk

A Weakly Supervised Bayesian Model for Violence Detection in Social Media

Elizabeth Cano **Yulan He**
School of Engineering and Applied Science
Aston University, UK
{a.cano_basave,y.he9}@aston.ac.uk

Kang Liu **Jun Zhao**
Institute of Automation
Chinese Academy of Sciences, China
{kliu,jzhao}@nlpr.ia.ac.cn

Abstract

Social streams have proven to be the most up-to-date and inclusive information on current events. In this paper we propose a novel probabilistic modelling framework, called violence detection model (VDM), which enables the identification of text containing violent content and extraction of violence-related topics over social media data. The proposed VDM model does not require any labeled corpora for training, instead, it only needs the incorporation of word prior knowledge which captures whether a word indicates violence or not. We propose a novel approach of deriving word prior knowledge using the relative entropy measurement of words based on the intuition that low entropy words are indicative of semantically coherent topics and therefore more informative, while high entropy words indicates words whose usage is more topical diverse and therefore less informative. Our proposed VDM model has been evaluated on the TREC Microblog 2011 dataset to identify topics related to violence. Experimental results show that deriving word priors using our proposed relative entropy method is more effective than the widely-used information gain method. Moreover, VDM gives higher violence classification results and produces more coherent violence-related topics compared to a few competitive baselines.

1 Introduction

Social media and in particular Twitter has proven to be a faster channel of communication when compared to traditional news media, as we have witnessed during events such as the Middle East revolutions and the 2011 Japan earthquake; acting as social sensors of real-time events (Sakaki et al., 2010). Therefore the identification of topics discussed in these channels could aid in different scenarios including violence detection and emergency response. In particular the task of classifying tweets as violence-related poses different challenges including: high topical diversity; irregular and ill-formed words; event-dependent vocabulary characterising violence-related content; and an evolving jargon emerging from violent events.

Indeed, machine learning methods for classification present difficulty on short texts (Phan et al., 2008). A large body of work has been proposed for the task of topic classification of Tweets (Milne and Witten., 2008; Gabrilovich and Markovitch, 2006; Genc et al., 2011; Muñoz García et al., 2011; Kasiviswanathan et al., 2011; Meij et al., 2012). Recent approaches have also been proposed (Michelson and Macskassy, 2010; Cano et al., 2013), to alleviate microposts sparsity by leveraging existing social knowledge sources (e.g Wikipedia). However, while the majority of these approaches rely on supervised classification techniques, others do not cater for the violence detection challenges. To the best of our knowledge very few have been devoted to violent content analysis of Twitter, and none has carried out deep violence-related topic analysis. Since violence-related events tend to occur during short to medium life spans, traditional classification methods which rely on labelled data can rapidly become outdated. Therefore in order to maintain tuned models it is necessary the continuous learning from social media in order to capture those features representing violent events. Indeed, the task of violence classification demands more efficient and flexible algorithms that can cope with rapidly evolving features. These observations have thus motivated us to apply unsupervised or weakly supervised approaches for domain-independent violence classification.

Another shortcoming of previous classification approaches is that they only focus on detecting the overall topical category of a document. However they do not perform an in-depth analysis to discover the latent topics and the associated document category.

International Joint Conference on Natural Language Processing, pages 109–117,
Nagoya, Japan, 14-18 October 2013.

When examining violence-related data, analysts are not only interested in the overall violence of one particular tweet but on the understanding of the type of emerging violence-related events. For example the word "killing" may have a violent-related orientation as in "mass killing" while it has a non-violent one in "killing time". Therefore, detecting topic and violence-relatedness simultaneously should serve as a critical function in helping analysts by providing more informative violence-related topic mining results.

In this paper, we introduce the Violence Detection Model (VDM), which focuses on document-level violence classification for general domains in conjunction with topic detection and violence-related topic analysis. The model extends the Latent Dircichlet Allocation (LDA) (Blei et al., 2003) by adding a document category (violent or non-violent) layer between the document and the topic layer. It is related to the joint sentiment-topic (JST) model for simultaneous sentiment and topic detection (Lin and He, 2009; Lin et al., 2012). However, while JST assumes the per-document sentiment-topic distributions, VDM only has a single document category-topic distribution shared across all the documents. This is because tweets are short compared to typical review documents and hence modelling per-tweet category-topic distribution could potentially generate less coherent topics. In VDM, we also assume that words are generated either from a category-specific topic distribution or from a general background model. This helps reducing the effects of background words and learn a model which better captures words concentrating around category-specific topics. As will be discussed later, VDM outperforms JST in both violence detection from tweets and topic coherence measurement. Furthermore, while JST incorporates word prior sentiment knowledge from existing sentiment lexicons, we propose a novel approach to derive word prior knowledge based on the relative entropy measurement of words.

We proceed with related work on topic classification on Twitter. Since the Bayesian model studied here is closely related to the LDA model, we also review existing approaches of incorporating supervised information into LDA training. We then present our proposed VDM model and describe a novel approach of deriving word priors using relative entropy from DBPedia[1] articles and tweets annotated using OpenCalais[2]. Following that, we present the dataset used in the paper and discuss experimental results obtained in comparison to a few baselines. Finally, we conclude the paper.

[1] http://dbpedia.org

[2] http://www.opencalais.com

2 Related Work

The task of detecting violent-related tweets can be viewed as a topical classification (TC) problem in which a tweet is labelled either as violent or non-violent related. Since the annotation of Twitter content is costly, some approaches have started to explore the incorporation of features extracted from external knowledge sources (KS) and the use of unsupervised or semi-supervised approaches to solve the TC problem. Since the model proposed in this paper makes use of both external KSs and topic models, we have divided the review of related work into approaches which rely on external KSs and approaches based on LDA model learning.

In the first case, Genc et al. (2011) proposed a latent semantic topic modelling approach, which mapped a tweet to the most similar Wikipedia[3] articles based on the tweets' lexical features. Song et al. (2011) mapped a tweet's terms to the most likely resources in the Probbase KS. These resources were used as additional features in a clustering algorithm which outperformed the simple bag of words approach. Munoz et al. (2011) proposed an unsupervised vector space model for assigning DBpedia URIs to tweets in Spanish. They used syntactical features derived from PoS (part-of-speech) tagging, extracting entities using the Sem4Tags tagger (Garcia-Silva et al., 2010) and assigning DBpedia URIs to those entities by considering the words appearing in the context of an entity inside the tweets. In contrast to these approaches, rather than labelling a tweet with KS URIs, we make use of DBpedia violence-related articles as one possible source of information from which prior lexicons can be derived.

Recently, Cano et al. (2013) proposed a supervised approach which makes use of the linked structure of multiple knowledge sources for the classification of Tweets, by incorporating semantic metagraphs into the feature space. However, in this study rather than extending the feature space with DBpedia derived features, we propose a strategy for characterising Violence related topics through the use of relative entropy, which filters out irrelevant word features. Moreover the proposed VDM model not only classifies documents as violent-related but also derives coherent category-topics (collection of words labelled as violent-related and non-violent related).

Our VDM model incorporates word prior knowledge into model learning. Here, we also review existing approaches for the incorporation of supervised information into LDA model learning. The supervised LDA (sLDA) (Blei and McAuliffe, 2008) uses empirical topic frequencies as a covariant for

[3] http://wikipedia.org

a regression on document labels such as movie ratings. The Dirichlet-multinomial regression (DMR) model (Mimno and McCallum, 2008) uses a log-linear prior on document-topic distributions that is a function of observed meta data of the document. Labeled LDA (Ramage et al., 2009) defines a one-to-one correspondence between LDA's latent topics and observed document labels and utilize a transformation matrix to modify Dirichlet priors. Partially Labeled LDA (PLDA) extends Labeled LDA to incorporate per-label latent topics (Ramage et al., 2011). The DF-LDA model (Andrzejewski et al., 2009) employs must-link and cannot-link constraints as Dirichlet Forest priors for LDA learning, but it suffers the scalability issue. Most recently, the aspect extraction model for sentiment analysis (Mukherjee and Liu, 2012) assumes that a seed set is given which consists of words together with their respective aspect category. Then depending on whether a word is a seed or non-seed word, a different route of multinomial distribution will be taken to emit the word. Our work was partially inspired by the previously proposed joint sentiment-topic model (JST) (Lin and He, 2009; Lin et al., 2012), which extracts topics grouped under different sentiments, relying only on domain-independent polarity word prior information.

While the afore-mentioned approaches assume the existence of either document label information or word prior knowledge, we propose to learn word prior knowledge using relative entropy from DBpedia and tweets annotated using OpenCalais. Moreover the proposed VDM model relies on the assumptions that the document category-topic distribution is shared across all documents in a corpus and words are generated either from a category-specific topic distribution or from a general background distribution. As we will discuss in section 5 these assumptions along with the proposed strategies for prior lexicon derivation show promising results outperforming various other topic models.

3 Violence Detection Model (VDM)

We propose a weakly-supervised violence detection model (VDM) here. In this model violence labels are associated with documents, under which topics are associated with violence labels and words are associated with both violence labels and topics. The graphical model of VDM is shown in Figure 1.

Assume a corpus of D documents denoted as $\mathcal{D} = \{\boldsymbol{d}_1, \boldsymbol{d}_2, .., \boldsymbol{d}_D\}$; where each document consists of a sequence of N_d words denoted by $\boldsymbol{d} = (w_1, w_2, .., w_{N_d})$; and each word in a document is an item from a vocabulary index of V different terms denoted by $1, 2, .., V$. We also assume that when an author writes a tweet message, she first decides whether

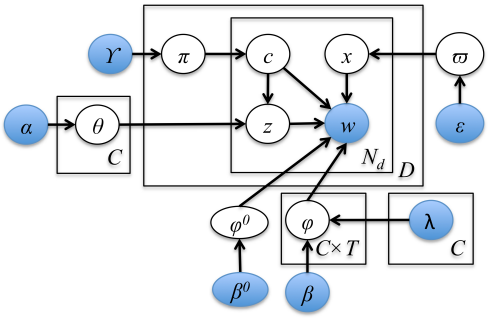


Figure 1: Violence detection model (VDM).

the tweet is violent-related or not. We use a category variable c to indicate violent-related topics or non-violent topics. If $c = 0$, the tweet is non-violent and the tweet topic is drawn from a general topic distribution θ_0. If $c = 1$, the tweet is violent-related and the tweet topic is drawn from a violent category specific topic distribution θ_1. Finally, each word of the tweet message is generated from either the background word distribution ϕ^0, or the multinomial word distribution for the violent-related topics $\phi_{c,z}$. The generative process of VDM is shown below.

- Draw $\omega \sim \text{Beta}(\epsilon), \varphi^0 \sim \text{Dirichlet}(\beta^0), \varphi \sim \text{Dirichlet}(\beta)$.
- For each tweet category $c = 1, ..., C$,
  - for each topic z under the tweet category c, draw $\theta_{cz} \sim \text{Dirichlet}(\alpha)$.
- For each document $m \in \{1..D\}$,
  - draw $\pi_m \sim \text{Dirichlet}(\gamma)$,
  - For each word $n \in \{1..N_d\}$ in document m,
    - draw $x_{m,n} \sim \text{Multinomial}(\omega)$;
    - if $x_{m,n} = 0$,
      - draw a word $w_{m,n} \sim \text{Multinomial}(\varphi^0)$;
    - if $x_{m,n} = 1$,
      - draw a tweet category label $c_{m,n} \sim \text{Multinomial}(\pi_m)$,
      - draw a topic $z_{m,n} \sim \text{Multinomial}(\theta_{c_{m,n}})$,
      - draw a word $w_{m,n} \sim \text{Multinomial}(\varphi_{c_{m,n},z_{m,n}})$.

We have a latent random variable x associated with each word token and acts as a switch. If $x = 0$, words are generated from a background distribution. If $x = 1$, words are sampled from the corpus-specific multinomial $\varphi_{c,z}$ decided by the tweet category label (non-violent or violent) c and the tweet topic z.

3.1 Model Inference

We use Collapsed Gibbs Sampling (Griffiths and Steyvers, 2004) to infer the parameters of the model and the latent violent categories and topics assignments for tweets, given observed data $\mathcal{D}$. Gibbs sampling is a Markov chain Monte Carlo method which

allows us to repeatedly sample from a Markov chain whose stationary distribution is the posterior of interest, switch variable x, category label c, and topic z here, from the distribution over that variable given the current values of all other variables and the data. Such samples can be used to empirically estimate the target distribution. Letting the index $t = (m, n)$ denote n^{th} word in document m and the subscript $-t$ denote a quantity that excludes data from n^{th} word position in document m, the conditional posterior for x_t is:

$$P(x_t = 0|\boldsymbol{x}_{-t}, \boldsymbol{c}, \boldsymbol{z}, \boldsymbol{w}, \Lambda) \propto \frac{\{N_m^0\}_{-t} + \epsilon}{\{N_m\}_{-t} + 2\epsilon} \times \frac{\{N_{w_t}^0\}_{-t} + \beta^0}{\sum_{w'}\{N_{w'}\}_{-t} + V\beta^0}, \quad (1)$$

where N_m^0 denotes the number of words in document m assigned to the background component, N_m is the total number of words in document m, $N_{w_t}^0$ is the number of times word w_t is sampled from the background distribution.

$$P(x_t = 1|\boldsymbol{x}_{-t}, \boldsymbol{c}, \boldsymbol{z}, \boldsymbol{w}, \Lambda) \propto \frac{\{N_m^s\}_{-t} + \epsilon}{\{N_m\}_{-t} + 2\epsilon} \times \frac{\{N_{w_t}^s\}_{-t} + \beta}{\sum_{w'}\{N_{w'}\}_{-t} + V\beta}, \quad (2)$$

where N_m^s denotes the number of words in document m sampled from the category-topic distributions, $N_{w_t}^s$ is the number of times word w_t is sampled from the category-topic specific distributions.

The conditional posterior for c_t and z_t is:

$$P(c_t = k, z_t = j|\boldsymbol{c}_{-t}, \boldsymbol{z}_{-t}, \boldsymbol{w}, \Lambda) \propto \frac{N_{d,k}^{-t} + \gamma}{N_d^{-t} + C\gamma} \cdot \frac{N_{d,k,j}^{-t} + \alpha_{k,j}}{N_{d,k}^{-t} + \sum_j \alpha_{k,j}} \cdot \frac{N_{k,j,w_t}^{-t} + \beta}{N_{k,j}^{-t} + V\beta}, \quad (3)$$

where $N_{d,k}$ is the number of times category label k has been assigned to some word tokens in document d, N_d is the total number of words in document d, $N_{d,k,j}$ is the number of times a word from document d has been associated with category label k and topic j, N_{k,j,w_t} is the number of times word w_t appeared in topic j and with category label k, and $N_{k,j}$ is the number of words assigned to topic j and category label k.

Once the assignments for all the latent variables are known, we can easily estimate the model parameters $\{\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{\varphi}, \boldsymbol{\varphi^0}, \boldsymbol{\omega}\}$. We set the symmetric prior $\epsilon = 0.5$, $\beta_0 = \beta = 0.01$, $\gamma = (0.05 \times L)/C$, where L is the average document length, C the is total number of category labels, and the value of 0.05 on average allocates 5% of probability mass for mixing. The asymmetric prior $\boldsymbol{\alpha}$ is learned directly from data using maximum-likelihood estimation (Minka, 2003) and updated every 40 iterations during the Gibbs sampling procedure. We run Gibbs sampler for 1000 iterations and stop the iteration once the log-likelihood of the training data converges under the learned model.

3.2 Deriving Model Priors through Relative Entropy

Detecting violence and extremism from text closely relates to sentiment and affect analysis. While sentiment analysis primarily deals with positive, negative, or neutral polarities, affect analysis aims to map text to much richer emotion dimensions such as joy, sadness, anger, hate, disgust, fear, etc. In the same way violence analysis maps violence polarity into violence words such as looting, revolution, war, drugs and non-violent polarity to background words such as today, happy, afternoon. However, as opposed to sentiment and affect prior lexicon derivation, the generation of violence prior lexicons pose different challenges. While sentiment and affect lexicon, rarely changes in time, words relevant to violence tend to be event dependent.

In this section we introduce a novel approach for deriving word priors from social media, which is based on the measurement of the relative entropy of a word in a corpus. Assume a source corpus consisting of N documents denoted as $\mathcal{SD} = \{\boldsymbol{sd}_1, \boldsymbol{sd}_2, ..., \boldsymbol{sd}_N\}$, where each document is labelled as not violent or violent. We define the following metrics:

1. **Corpus Word Entropy**: The entropy of word w in corpus $\mathcal{SD}$ is measured as follows:

$$E_{SD}(w) = -\sum_{i=1}^{N} p(w|\boldsymbol{sd}_i) \log p(w|\boldsymbol{sd}_i), \quad (4)$$

where $p(w|\boldsymbol{sd}_i)$ denotes the probability of word w given the document $\boldsymbol{sd}_i$ and N the total number of documents. $E_{SD}(w)$ captures the dispersion of the usage of word w in the corpus . Our intuition is that low entropy words are indicative of semantically coherent topics and therefore more informative, while high entropy words indicates words whose usage is more topical diverse and therefore less informative.

2. **Class Word Entropy**: The entropy of word w given the class label c is defined as follows:

$$E_{CWE}(w, c) = -\sum_{i=1}^{N} p(w|\boldsymbol{sd}_i^c) \log p(w|\boldsymbol{sd}_i^c), \quad (5)$$

where C denotes the number of classes (in our case violent and non-violent) and $p(w|\boldsymbol{sd}_i^c)$ denotes the probability of word w given the document $\boldsymbol{sd}_i$ in class c. In contrast to the general E_{SD}, the class word entropy characterises the usage of a word in a particular document class.

3. **Relative Word Entropy (RWE)**: In order to compare the word entropy used on documents in different categories, we measure the word relative entropy as follows:

$$RWE(w,c) = \frac{E_{CWE}(w,c)}{E_{SD}(w)} \quad (6)$$

The RWE provides information on the relative importance of that word to a given document class.

After deriving the RWE of each word given a class (i.e violent or non-violent), we sorted words based on their RWE values in ascending order. Since our intuition is that lower entropy levels are more indicative of semantically coherent topics we choose the top K words of each class. We then built a matrix $\boldsymbol{f}$ of size $K \times C$, where C is the total number of document classes or category labels. The kcth entry stores the probability that feature k is assigned with category label c. The matrix $\boldsymbol{f}$ essentially captures word prior knowledge and can be used to modify the Dirichlet prior $\boldsymbol{\beta}$ of category-topic-word distributions. We initialize each element of the matrix $\boldsymbol{\beta}$ of size $C \times T \times V$ to 0.01 and then perform element-wise multiplication between $\boldsymbol{\beta}$ and $\boldsymbol{f}$ with the topic dimension ignored.

4 Experimental Setup

4.1 Dataset Description

The experimental setup consists of three stages: 1) derivation of word prior lexicon; 2) training of VDM and baselines; and 3) testing. For the first stage, we explored three different ways to construct a labelled document corpora for deriving prior lexicons. The first one is based on a Twitter corpus labelled using OpenCalais. This corpus comprises over 1 million tweets collected over a period of two months starting from November 2010. In order to build the Twitter-based violent dataset for deriving priors, we extracted tweets labelled as "War & Conflict" and considered them as violent annotations, while for the non-violent annotations we considered tweets annotated with labels other than this one (e.g. Education, Sports). We denote this dataset as **TW**. It is worth noting that the annotated results generated by OpenCalais are very noisy. We have evaluated OpenCalais on our manually annotated test set and only obtained an F-measure of 38%. Nevertheless, as will be seen later, word prior knowledge extracted from such noisy annotated tweets data is still very helpful in learning the VDM model for violence detection from tweets.

The second dataset for deriving priors is based on DBpedia which is a knowledge source derived from Wikipedia. The latest version of DBpedia consists of over 1.8 million resources, which have been classified into 740 thousand Wikipedia categories, and over 18 million YAGO[4] categories. For constructing the violence related corpus we queried DBpedia for all articles belonging to categories and subcategories under the "violence" category, from which we kept their abstract as the document content. After removing those categories with less than 1000 articles, we obtained a set of 28 categories all related to violence. The resulting set of articles represented the violent set while for the non-violent rather than using non-violent related articles from DBpedia we opted for using the collection of Tweets from **TW** annotated as non-violent by OpenCalais. This decision was made in order to balance differences across the DBpedia and Twitter lexicons. This resulting dataset is referred to as **DB**.

Since the average word per article abstract in DBpedia exceeds the one of tweets, we decided to build a third dataset where the violent DBpedia documents resemble tweets in their size. In order to do so, we took into account that the average number of words per tweet in **TW** before preprocessing is 9.6. Then from each violent document in the **DB** dataset, we generated tweet size documents by chunking the abstracts into 9 or less words. We then combine the chunked documents from **DB** with **TW** and refer to the final dataset as **DCH**.

These datasets were used for deriving priors for the first stage. For the second stage, we built a training set of tweets derived from the TREC Microblog 2011 corpus[5], which comprises over 16 million tweets sampled over a two week period (January 23rd to February 8th, 2011). This time period includes 49 different events including violence-related ones such as Egyptian revolution, and Moscow airport bombing, and non-violence related such as the Super Bowl seating fiasco. We sampled a subset of 10,581 tweets as our training set and manually annotated another 1,759 tweets as our test set. Details about the statistics of the training and testing datasets are presented in Table 1 under the label "Main Dataset".

We preprocessed the described datasets by first removing: punctuation, numbers, non-alphabet characters, stop words, user mentions, links and hashtags. We then performed Lovins stemming in order to reduce the vocabulary size. Finally to address the issue of data sparseness, we removed words with a frequency lower than 5.

4.2 Deriving Model Priors

We derive word prior knowledge from the three datasets mentioned above, namely **TW**, **DB** and **DCH**; applying the relative word entropy (RWE)

[4] http://www.mpi-inf.mpg.de/yago-naga/yago/

[5] http://trec.nist.gov/data/tweets/

	Datasets for Priors		
	TW	DB	DCH
Vio	10,432	4,082	32,174
Non-Vio	11,411	11,411	11,411
	Main Dataset		
	Training Set		Testing Set
Vio	10,581		759
Non-Vio			1000

Table 1: Document statistics of the datasets used for deriving prior lexicons and for training and testing the proposed model and baselines.

approach introduced in section 3.2 for word prior lexicon generation. For comparison purposes, we also employ the widely-used information gain (IG) method to select highly discriminative words under each class from the datasets. Table 2 presents the word statistics of the prior lexicons generated using these two different methods[6]. It worth noting that **DB** consists of 4,082 violent-related documents (DBpedia abstracts) and 11,411 non-violent documents (non-violent tweets). Since the average word per abstract is much larger in size than the one of a tweet, having a very low number of non-violent features selected using IG is expected as the violence class is over represented per violent document. This is the reason why we built another dataset by chunking the DBpedia abstracts to produce tweet-size documents (**DCH**). Having a balanced number of words per document in both violent and non-violent categories leads to more balanced priors, as shown in Table 2, where the number of non-violent features increased from 99 (in **DB**) to 1,345 (in **DCH**) using IG.

	IG			RWE		
	TW	DB	DCH	TW	DB	DCH
Vio	1,249	2,899	1,612	875	3,388	3,786
Non-Vio	1,749	99	1,345	2,595	879	2,438

Table 2: Statistics of the word prior lexicons.

4.3 Baselines

For comparison purposes, we have tested the following baselines:

Learned from labelled features. The word prior knowledge can be used as labelled feature constraints which can be incorporated into a MaxEnt classifier training with Generalized Expectation (GE) constraints (Druck et al., 2008) or Posterior Regularization (PR) (Ganchev et al., 2010). We use the implementation provided in MALLET with default parameter configurations for our experiments and refer these two methods as *ME-GE* and *ME-PR* respectively.

JST. If we set the number of sentiment classes to 2 (violent or non-violent), then we can learn the Joint Sentiment-Topic (JST) model from data with the word prior knowledge incorporated in a similar way as the VDM model.

PLDA. The Partially-Labeled LDA (PLDA) (Ramage et al., 2011) model assumes that some document labels are observed and models per-label latent topics. It is somewhat similar to JST and VDM except that supervised information is incorporated at the document level rather than at the word level. The training set is labelled as violent or non-violent using OpenCalais. Such pseudo document labels are then incorporated into PLDA for training.

The hyperparameters of PLDA and JST are set to be the same as those for VDM.

5 Experimental Results

In this section we compare the overall classification performance of VDM and a set of proposed baselines. We performed a series of experiments to investigate the impact of the prior derivation strategies (RWE and IG) on classification performance, using the six prior lexicons introduced in Section 4.2. Some of the research questions addressed in this section are as follows: Do lexicons built from DBpedia contain useful features which can be applied for the violence classification of Tweets?; If so, to what extent these lexicons help the classification task?. We also present the overall evaluation of the proposed VDM against the proposed baselines based on the semantic coherence of the generated topics. All the experiments reported here were conducted using a 5 fold 3 trial setting.

5.1 Violence Classification Results vs. Different Word Priors

Table 3 compares the results obtained for violence classification for the proposed VDM model against the baselines, using prior lexicons derived with the proposed RWE strategy and the IG baseline approach. We can observe that although both ME-GE and ME-PR present a very high precision for word priors obtained from **TW** regardless using either IG or RWE, they also present a very low recall. This indicates that although the documents labelled as "violent" with these models were correctly identified, much of the rest of the violent documents in the testing set remained unidentified. We can also observe that the best results in terms of F-measure were obtained for the VDM model using the word priors derived from **TW** using RWE, which significantly outper-

[6]While the number of words selected for IG was set to 3000, the criteria for selecting the top K words in the RWE approach was based on taking the highest coherent level of entropy containing more that 5 words. Then from the sorted list of words we selected those whose entropy was smaller than this level.

	Prior	ME-GE			ME-PR			JST			VDM		
		P	R	$F1$	P	R	$F1$	P	R	$F1$	P	R	$F1$
IG	TW	0.7737	0.0337	0.0646	0.6300	0.1034	0.1777	0.6939	0.9362	0.7969	0.75	0.9288	0.8297
	DB	0.4604	0.9704	0.6245	0.5634	0.6955	0.4773	0.6493	0.9228	0.7622•	0.6455	0.9141	0.7566
	DCH	0.4862	0.2447	0.3255	0.5680	0.2949	0.3274	0.7113	0.9291	0.8057	0.7575	0.92	0.8309
RWE	TW	0.7100	0.1342	0.2249	0.9125	0.0373	0.0717	0.7235	0.9296	0.8136	0.8258	0.8919	**0.8575**
	DB	0.4958	0.1844	0.2686	0.5303	0.0540	0.0981	0.6882	0.9421	0.7952	0.7024	0.9212	0.7969
	DCH	0.5161	0.1731	0.2588	0.8485	0.0091	0.0179	0.73	0.9351	0.8199	0.8189	0.8804	0.8484

Table 3: The performance of the classifiers using prior features derived from TW, DB and DCH ($dbp + tw$). The number of topics is set to 5 for JST and VDM. The values highlighted in bold corresponds to the best results obtained in F-measure, while the shaded cells indicate the best results in F-measure for each scenario. Blank notes denotes that the F-measure of VDM significantly outperforms the baselines while • denotes JST outperforms VDM. Significance levels: p-value < 0.01

forms the baseline models (t-test with $\alpha < 0.01$). To compare VDM against JST, we varied the topics $T \in \{1, 5, 10, 15, 20, 25, 30\}$ and our significance test results revealed that VDM outperforms JST significantly (t-test with $\alpha < 0.01$) over all the topic settings except for the JST using **DB** lexicon priors.

When comparing the effectiveness of the use of DBpedia as a source of prior lexicon, we can observe that the use of the full articles' abstracts in the derivation of the prior lexicons **DB** did not present an improvement over the models based on Twitter derived lexicons (**TW**). However, the strategy of chunking DBpedia articles' abstracts into tweet size documents (**DCH**), did help in boosting the overall F-measure in JST (t-test with $\alpha < 0.05$). In the case of VDM, the use of **DCH** achieved an F-measure very close to the one obtained using Twitter prior lexicons (**TW**).

When comparing the effectiveness of the proposed RWE strategy against the IG baseline for deriving prior lexicons, we can observe that RWE consistently outperformed in F-measure for the JST and VDM models on all the three prior lexicon scenarios with the improvement ranging between 1-4% although it fails to boost F-measure on both ME-GE and ME-PR.

In the subsequent experiments, we incorporated word prior knowledge extracted from **TW** using our proposed RWE method.

5.2 Varying Number of Topics

We compare the violence classification accuracy of our proposed VDM model against PLDA and JST with different topic number settings. It can be observed from Figure 2 that with single topic setting, all the three models give a similar violence classification results. However, when increasing the number of topics, PLDA performs much worse than both JST and VDM with the violence classification accuracy stabilising around 60%. In PLDA, document labels of the training set were obtained using OpenCalais. As mentioned in Section 4.1, OpenCalais gave an F-measure of 38% for violence classification on the test set. Hence document labels of the training set are not reliable. This explains the low classification accuracy of PLDA.

VDM gives fairly stable violence classification results across different topic numbers. The violence classification accuracy using JST attains the best with single topic and drops slightly with the increasing number of topics. This is because JST assumes the per-tweet category-topic distribution and potentially generates less coherent topics which affects the violence classification accuracy.

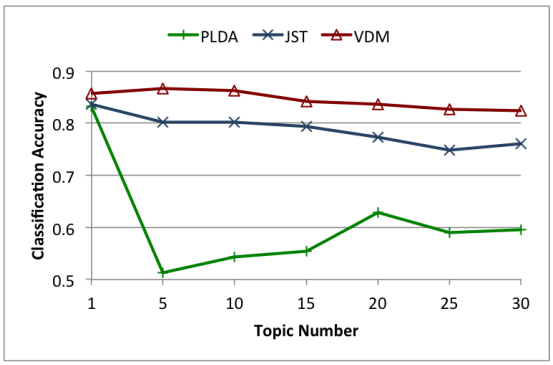


Figure 2: Violence classification Accuracy versus different topic numbers.

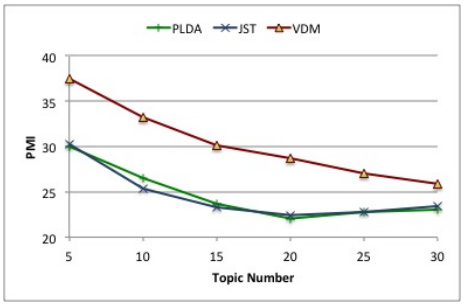
5.3 Topic Extraction

Table 4 presents two topic examples of violent and non-violent topics generated by VDM, JST and PLDA. We can observe that the topics revealed by VDM are representative of some of the events appearing during January/February 2011. For example, T1 gives an insight on the spreading of the Middle East Arab revolution, while T2 provides information regarding the Moscow airport bombing. For the case of non-violent topics, VDM revealed topics which appeared to be less semantically coherent than those of violent topics. However when reading the non-violent VDM T1, it gives an insight of the super bowl game related to the Jets. When checking the topics revealed

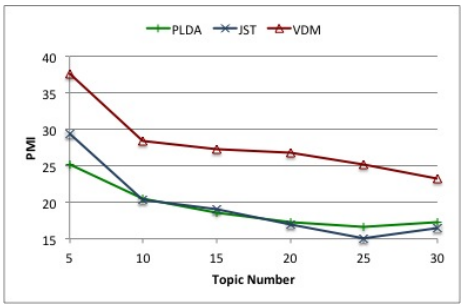
VDM				JST				PLDA			
Violent		Non-Violent		Violent		Non-Violent		Violent		Non-Violent	
T1	T2	T1	T2	T1	T2	T1	T2	T1	T2	T1	T2
middle	crash	bowl	people	middle	crash	game	day	kill	game	crash	wow
east	kill	game	hate	government	nat	win	good	moscow	win	polic	cut
give	moscow	win	give	police	museum	jets	free	bomb	jets	drug	block
power	bomb	jets	damn	revolution	moscow	bowl	people	airport	watch	protester	arm
idea	airport	fan	shit	world	loot	fan	thing	leave	today	arrest	till
government	tweets	watch	miss	arm	report	reason	work	islam	play	car	officer
live	thought	today	fuck	streets	bomb	go	hope	injure	car	people	nat
time	injure	gone	hah	day	airport	damn	life	crash	fan	kill	fire
fall	arrest	damn	close	watch	kill	injure	today	report	damn	top	support
spread	dead	car	guy	live	morn	play	hah	victim	hate	part	london
upris	world	friends	sense	support	secure	run	back	terror	best	show	american

Table 4: Topic examples extracted under Violent and Non-Violent Labels for topic setting of 30 topics.

by JST, we can observe that although words seem to be semantically coherent for both violent and non-violent topics, there are words which belong to different violent events. For example the JST violent T2 mixes the Moscow bombing event with the Egyptian protesters Museum attack event. When checking the topics produced by PLDA we can see that it fails to correctly characterise violen and no-violent topics, since PLDA T2 should have been clearly classified as non-violent and the non-violent PLDA T1 as violent. Moreover in the violent PLDA T1 topic which presents violent related words, we can empirically identify more than one event involved.



(a) Violent topics.



(b) Non-violent topics.

Figure 3: Topic coherence measurement based on PMI. A larger PMI value indicates a better model.

In order to measure the semantic topical coherence of VDM and the proposed baselines, we made use of the Pointwise Mutual Information(PMI) metric proposed in (Newman et al., 2010). PMI is an automatic topic coherence evaluation which has been found to correspond well with human judgements on topic coherence. In particular, a coherent topic should only contain semantically related words and hence any pair of the top words from the same topic should have a large PMI value. For each topic, we compute its PMI by averaging over the PMI of all the word pairs extracted from the top 10 topic words. Figure 3 shows the PMI values of topics extracted under the violence and non-violence classes with the topic numbers varying between 5 and 30. It can be observed that JST and PLDA give similar PMI results. However, VDM outperforms both by a large margin.

6 Conclusions and Future Work

In this paper, we have proposed a novel violence detection model (VDM), which enables the identification of text containing violent content and extraction of violence-related topics over social media data. VDM learning requires the incorporation of word prior knowledge which captures whether a word indicates violence or not. We propose a novel approach of deriving word prior knowledge using the measurement of relative entropy of words (RWE). Extensive experiments on the tweets data sampled from the TREC Microblog 2011 dataset show that our proposed RWE is more effective in deriving word prior knowledge compared to information gain. Moreover, the VDM model gives significantly better violence classification results compared to a few competitive baselines. It also extracts more coherent topics.

In future work, we intend to explore online learning strategies for VDM to adaptively update its parameters so that it can be used for violence detection from social streaming data in real-time.

Acknowledgments

This work was partially supported by the EPRSC and DSTL under the grant EP/J020427/1 and the Visiting Fellowship funded by the National Laboratory of Pattern Recognition, Insitute of Automation, Chinese Academy of Sciences.

References

David Andrzejewski, Xiaojin Zhu, and Mark Craven. 2009. Incorporating domain knowledge into topic modeling via dirichlet forest priors. In *ICML*, pages 25–32.

D.M. Blei and J. McAuliffe. 2008. Supervised topic models. In *NIPS*, 20:121–128.

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

A. E. Cano, A. Varga, F. Ciravegna, and Y. He. 2013. Harnessing linked knowledge source for topic classification in social media. In *Proceeding of the 24th ACM Conference on Hypertext and Social Media (Hypertext)*.

G. Druck, G. Mann, and A. McCallum. 2008. Learning from labeled features using generalized expectation criteria. In *SIGIR*, pages 595–602.

Evgeniy Gabrilovich and Shaul Markovitch. 2006. Overcoming the brittleness bottleneck using Wikipedia: enhancing text categorization with encyclopedic knowledge. In *AAAI*.

Kuzman Ganchev, João Graça, Jennifer Gillenwater, and Ben Taskar. 2010. Posterior regularization for structured latent variable models. *Journal of Machine Learning Research*, 11:2001–2049.

A. Garcia-Silva, Oscar Corcho, and J. Gracia. 2010. Associating semantics to multilingual tags in folksonomies.

Yegin Genc, Yasuaki Sakamoto, and Jeffrey V. Nickerson. 2011. Discovering context: classifying tweets through a semantic transform based on wikipedia. In *Proceedings of the 6th international conference on Foundations of augmented cognition: directing the future of adaptive systems*, pages 484–492, Berlin, Heidelberg. Springer-Verlag.

T.L. Griffiths and M. Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101(Suppl 1):5228?5235.

Shiva Prasad Kasiviswanathan, Prem Melville, Arindam Banerjee, and Vikas Sindhwani. 2011. Emerging topic detection using dictionary learning. In *CIKM*, pages 745–754, New York, NY, USA. ACM.

C. Lin and Y. He. 2009. Joint sentiment/topic model for sentiment analysis. In *CIKM*, pages 375–384.

C. Lin, Y. He, R. Everson, and S. Rüger. 2012. Weakly-Supervised Joint Sentiment-Topic Detection from Text. *IEEE Transactions on Knowledge and Data Engineering*, 24(6):1134–1145.

Edgar Meij, Wouter Weerkamp, and Maarten de Rijke. 2012. Adding semantics to microblog posts. In *WSDM*, pages 563–572.

Matthew Michelson and Sofus A. Macskassy. 2010. Discovering users' topics of interest on twitter: a first look. In *Proceedings of the fourth workshop on Analytics for noisy unstructured text data*, New York, NY, USA.

D. Milne and I. H. Witten., editors. 2008. *Learning to link with Wikipedia.*

D. Mimno and A. McCallum. 2008. Topic models conditioned on arbitrary features with dirichlet-multinomial regression. In *UAI*.

T. Minka. 2003. Estimating a Dirichlet distribution. Technical report.

Óscar Muñoz García, Andrés García-Silva, Óscar Corcho, Manuel de la Higuera Hernández, and Carlos Navarro. 2011. Identifying Topics in Social Media Posts using DBpedia. In *Proceedings of the NEM Summit*, pages 81–86, September.

Arjun Mukherjee and Bing Liu. 2012. Aspect extraction through semi-supervised modeling. In *ACL*, pages 339–348.

David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic evaluation of topic coherence. In *NAACL*, pages 100–108.

X. H. Phan, L. M. Nguyen, and S. Horiguchi. 2008. Learning to classify short and sparse text and web with hidden topics from large-scale data collections. In *WWW*.

D. Ramage, D. Hall, R. Nallapati, and C.D. Manning. 2009. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *EMNLP*, pages 248–256.

D. Ramage, C.D. Manning, and S. Dumais. 2011. Partially labeled topic models for interpretable text mining. In *KDD*.

Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. 2010. Earthquake shakes twitter users: real-time event detection by social sensors. In *WWW*, pages 851–860.

Yangqiu Song, Haixun Wang, Zhongyuan Wang, Hongsong Li, and Weizhu Chen. 2011. Short text conceptualization using a probabilistic knowledgebase. In *IJCAI*, pages 2330–2336.