Cross-Media Sentiment Classification and Application to Box-Office Forecasting

Élie Guàrdia-Sebaoun
LIP6, Paris France
elie.guardia-sebaoun@lip6.fr

Abdelhalim Rafrafi
LIP6, Paris France
abdelhalim.rafrafi@lip6.fr

Vincent Guigue
LIP6, Paris France
vincent.guigue@lip6.fr

Patrick Gallinari
LIP6, Paris France
patrick.gallinari@lip6.fr

ABSTRACT

This article aims at demonstrating the interest of opinion mining on Twitter data for the box-office prediction. Whilst most approaches in box-office forecasting focus on expert knowledge (actor celebrity, film budget...), or more recently on Twitter volumetric features, we want to show that the tweet's content is also important to make an efficient decision. Firstly we focus on the cross-media sentiment classification task, by studying the impact different algorithms and data sources have on the accuracy of sentiment classification on Twitter. Secondly, models allow us to to build high level sentiment features for the box-office forecasting problem. We demonstrate the interest of opinion mining derived features for this second task.

Keywords

Sentiment analysis, Cross-media adaptation, Box-office prediction, SVM, Least-square regression

1. INTRODUCTION

User generated contents and opinion mining techniques have received a significant attention during the last decade. Numerous economic issues such as survey, e-reputation managing and buzz detection are closely linked to them. Twitter and micro-blogging platforms represent a great source of data: posts are time stamped and linked to some users while the diffusion of the information is quick. Last but not least, those new sources gather millions of people' opinions and comments. Although the information is very noisy, the majority of topics which appear in Twitter and each event is almost reported in real time. Our aim is two folds: we want to predict sentiments in tweets content as well as to build high level sentiment features to forecast the box-office.

As Twitter data is unlabeled, it is thus uneasy tu use it in opinion mining applications. Our solution lies in the feedback tools available on most modern websites: web 2.0 users can now post some comments about any newspaper articles, e-commerce products or blog contributions. The key point lies in the star rating that often goes along those comments. Indeed, using those labels allow us to train supervised algorithms and to build efficient sentiment classifiers [27, 26].

on the one hand, Twitter provides us with dynamic and easily retrievable data but no label. On the other hand, we can collect (or simply download) huge review datasets with explicit sentiment labeling. We offer to analyze the cross-media transfer performance of machine learning algorithms on Twitter. In the sentiment classification, domain adaptation is known to be difficult since opinion markers differ from one topic to another [3]. The adaptation becomes even more complex in the cross-media context: the Twitter vocabulary relies on a lot of abbreviations and particular expression that are not used in reviews and blogs.

This adaptation task has been considered previously in [21] and it has been shown that review based models are efficient on Twitter, especially if a sufficient training dataset is used. [21] also studies fusion strategies with multiple mono-domain classifiers but they conclude that it is more efficient to learn roughly on all the available data. We set a close setting but our models are learnt on larger sets. We use existing collections from Amazon [3, 13], DBLP [18] and TripAdvisor [34] that are respectively made of 300k, 5800k, 25k and 25k reviews. Then, we evaluate the accuracy on a Twitter Golden Standard manually labeled [4]. Different approaches are compared, some are basic techniques consisting in learning on one dataset and testing on another. Other are advanced ones, relying on explicit transfer model [3, 5] [1].

We use those models in a second series of experiments in order to try and forecast the box office of 32 films as in [6, 1]. The Twitter dataset [4] is made of two parts: a labeled Golden Standard and a larger unlabeled set. The latter corresponds to the results of film queries on a 6 months period. We offer to measure the impact of different sentiment features on the forecast accuracy. Given some sentiment analysis models, we compute some statistics for each film and we optimize a regression problem. We demonstrate the interest of this new piece of information in comparison with classical

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
OAIR'13 May 22-24, 2013, Lisbon, Portugal
Copyright 2013 CID 978-2-905450-09-8 .

[1] Transfer models are learnt on both review set and twitter set: the latter is also manually annotated, it comes from [30].

volumetric approaches (based on tweet counting). Our new high level sentiment features dramatically increase the box office forecasting accuracy.

Related works are described in section 2, all details about our models are provided in section 3. Our results regarding cross-media sentiment classification are analyzed in section 4 and those related to box-office prediction are given in section 5

2. RELATED WORK

This section is divided in three parts, we first introduce sentiment analysis, then we give an overview of the different transfer learning techniques and finally, we focus on the box-office forecasting task.

2.1 Sentiment Analysis

Sentiment Analysis (SA) consists in classifying texts with respect to their polarities (positive, negative) [27] or to their subjectivities (objective or subjective) [25, 17]. SA has been a topic of interest for more than a decade but it became even more important with the spread of blogs [22] and micro-blogging platforms [28]. Although the task is the same, developed techniques are data-dependent. The word distribution on Twitter is quite different from the one used in reviews. Twitter contains a lots of abbreviations, whereas reviews tend to be more elaborated. Blogs rely on a much longer text and higher level lexicon.

It has been shown that tweet polarity is sometimes easy to retrieve, due to the use of a universal shortened language [23]. On the contrary, blog posts, being longer and better structured texts, often contain many opposite opinions, which make them harder to analyze [21].

2.2 Transfer Learning

Over the last few years, one of the biggest issues in SA was to learn a classifier efficient on a new target from existing labeled training sets. It requires a powerful domain-adaptation process and many proposals have been made [32, 3, 24, 19, 5]. The cross-media transfer task is even more challenging and more recent [21].

As each domain has its own word distribution, the classical i.i.d assumption does not hold in multi-domain classification. The adaptation consists in generalizing the different domains using various techniques such as *regularization*, *semantic learning* or *explicit alignment*.

Regularization.

In text categorization, most classifiers are linear and rely on a bag-of-words (BoW) representation. It has been shown that this setting is efficient for sentiment classification [26]. Thus, the weight optimization procedure may be regularized to improve generalization (and adaptation). Some dedicated framework have been suggested for multi-domain sentiment analysis [8, 29, 9].

Explicit alignment.

Another way to adapt the models is to search for some *pivot* words in order to align the lexicon distributions [3, 24]. The idea is to use matrix factorization techniques to characterize the behavior of every word independently from their domain. In concrete terms, some new domain-independent features are built in a pre-training unsupervised stage, leading to learn the classifier over both classical and new features.

Semantic Learning.

The adaptation problem can also be solved by learning a general topic-independent semantic. In this approach, words are mapped onto a continuous space and the classifier is learnt in this space. Early works relied on PLSA derivatives [7, 16]; then some solutions based on LDA have been proposed [10] and the trend goes now to neural networks architectures, using auto-encoders [11], convolutional neural networks [2] and recursive neural networks [18].

2.2.1 Multi-Domain Settings

Two different approaches have been offered to deal with domain adaptation. Given a target set to classify, either one can focus on a single domain to train the classifier [3], or on gathring multiple sources to be used as a new learning set.

If we can question the suitability of the first approach on real data, [19] demonstrates the theoretical benefits of multi-sources domain adaptation, by considering the target as a mixture of source domains and the possibility to identify the contribution of any source. These results are nevertheless empirically contested by [35]: the authors introduce the notion of leave-one-out at a dataset level (*i.e.* using all the datasets but one for learning and the last one as the target) but they do not observe any benefits. Some newer experiments [8] based on slightly larger datasets corroborate *Mansour et al.*'s theory.

2.2.2 Cross-Media Adaptation

[21] offers a study about cross-media adaptation between blogs, reviews and micro-blogs, using four models : the single-media model (classifier learned on a single media and tested on the others), the double-media model, the three-media model (classifier learned on the three kind of sources, excluding the target domain[2]) and the three-media voting model (classifier learned on each stream + majority voting on each document).

The authors came to two conclusions. firstly, the three-source model outperforms every other models, even the three-source voting one, and secondly, if the transfer from reviews and blogs to Twitter fares quite well, the converse is false.

2.3 Forecasting Application

Everyday, thousands of people give their opinion on Twitter and other social networks about movies, brands and political parties. They provide a glimpse of what might be the overall word of mouth of the population. Unlike chat, these data can be analyzed and discussed, allowing us to follow the path of a whirlwind, or the propagation of an epidemic. Why not going further and try to forecast events such as market fluctuation [31], or video game sales [20]?

Unfortunately, the data is noisy and constantly evolving making it hard to process. The majority of the above suggested approaches focus on low level feature such as tweet counting [6, 31, 33]. They also enhance the representation using high level features inherent to the subject, such as - in the case of [6] - casting, marketing budget, first week-end gross, number of theatres or movie genre-. Recently, some text based approaches were suggested, as in [15].

[2] In their setting, many domains are available for each media

3. MODELS & SETTINGS

This section is divided in two parts: firstly, we describe the different sentiment classifiers that have been used in our experiment. Secondly, we give some details about the forecasting models which are based on least square regression.

3.1 Features

We are using a basic unigram representation for texts (classical bags of words) combined with a presence coding as recommended in [26]. Thus, each document becomes a vector $\mathbf{x} \in \mathbb{R}^d$ with $x_j \in \{0, 1\}$ and d being the size of the dictionary. No particular pre-processing is applied, we just keep the most frequent 5000 words as it is usually done [3] (namely $d = 5000$).

3.2 Sentiment Classifiers

Cross-media sentiment classification is a difficult task since the vocabulary differs between Twitter and reviews or blogs. Initially, we do not tackle this problem explicitly: based on previous experiments [21], we offer to focus on basic classifiers that can be learnt over large annotated corpus. The idea is to make up for the vocabulary difference by using more training data and thus maximizing the chance to meet more expressions. The positive influence of large learning sets on sentiment classification has been stated clearly in [2].

As a consequence, we focus on an efficient and scalable algorithm: Support Vector Machines. We choose the implementation SVMlight [14] and we keep all default settings: linear model and default regularization parameter.

The resulting linear classifier relies on a weight vector $\mathbf{w} \in \mathbb{R}^d$, with $d = 5000$. The decision function for a document $\mathbf{x}$ is of the form $f(\mathbf{x}) = \langle \mathbf{x}, \mathbf{w} \rangle = \sum_{j=1}^{d} x_j w_j$. In the training set, every documents are associated to a label $y \in \{-1, 1\}$ which respectively stands for negative and positive label. As recommended in [26], we use the standard conversion from the star rating: documents with one & two stars are marked as negative, documents with four & five stars are marked as positive and documents with three stars are removed (the opinion associated to such rating is ambiguous and varies from one person to another). Do notice how each coefficient w_j is linked to a word j of the dictionary and can be interpreted: a positive value means that documents which contains this word will be drawn to the class +1. The bigger w_j is, the more influence it has on the final classification. On a labeled corpus of N documents, the SVM learning algorithm consists in optimizing the following problem:

$$\mathbf{w}^{\star} = \arg\min_{\mathbf{w}} \|\mathbf{w}\|^2 + C \sum_{i=1}^{N} \max(0, 1 - f(\mathbf{x}_i) y_i) \qquad (1)$$

where C denotes the regularization tradeoff between the classification accuracy on the training set and the generalization ability.

3.3 Transfer models

We are using two explicit transfer models to improve our sentiment classifiers on tweets. *Structural Correspondence Learning* [3] is the first efficient solution that has been proposed for domain adaptation in sentiment classification and *Frustratingly Easy Domain Adaptation* [5] is an fast and scalable solution that proved its efficiency on many applications.

Structural Correspondence Learning (SCL) [3].

SCL consists in a three steps strategy: firstly, np pivot words are extracted and it is assumed that they will have the same behavior in the source and target domains. Secondly, np linear classifiers $\mathbf{w} \in \mathbb{R}^d$ are learnt (where d is the dictionary size): they predict if a pivot is in a given document or not. Thirdly, all weight vectors $\mathbf{w}$ are concatenated, a matrix factorization is then applied on the resulting matrix (SVD) so as to build a projection matrix $\Theta \in \mathbb{R}^{np \times d}$. Each review is then described using both a standard bag of words and some transfer features obtained by projecting the review using Θ.

In our experiment, $np = 100$ and we take as sentiment pivots the most frequent words in the source data and in the Twitter validation corpus [30]. In order to focus on sentiment markers, we actually search for the np most frequent words that appear in the sentiment lexicon [12]. The $\mathbf{w}$ are learned using svmlight [14]. As a consequence, 100 new features are added for each review.

Frustratingly Easy Domain Adaptation (FEDA) [5].

The idea is to extend the review representation using $n + 1$ replications of the dictionary (one general representation plus one replication for each source/target, n is the number of sources). In this article, we use a triple representation: general/reviews/tweets. A review $\mathbf{x}$ becomes $\mathbf{x}_e = [\mathbf{x}\ \mathbf{x}\ 0]$ and a tweet is represented as $\mathbf{x}_e = [\mathbf{x}\ 0\ \mathbf{x}]$. The approach is very efficient and easy to implement but requires some labeled data from the target domain (i.e. some labeled tweets). The interest of the separated description resides in the automated balancing: even if the number of reviews is much larger than the number of labeled tweets, all discriminant informations from both part will be extracted efficiently.

A SVM is then trained on the new set $\{(\mathbf{x}_{e,i}, y_i)\}_{i=1,\dots,N}$. We still use SVMlight implementation.

3.4 High Level Features & Forecasting Model

In our second series of experiments, we are trying to predict the box office of 32 films using an unlabeled Twitter dataset [4]. We are using a 2-steps process: we extract d_r numerical features for each film so as to build a vector $\mathbf{x}_r$. Each film $\mathbf{x}_r$ is associated to a box office y_r and we then learn a linear predictor $f(\mathbf{x}_r) = \langle \mathbf{x}_r, \mathbf{w}_r \rangle$ that approximates y_r, minimizing the regularized least square criterion:

$$\mathbf{w}_r^{\star} = \arg\min_{\mathbf{w}_r} \sum_{i=1}^{N_f} (f(\mathbf{x}_{r,i}) - y_{r,i})^2 + \lambda \|\mathbf{w}_r\|^2, N_f = 32 \qquad (2)$$

We compute the following features for each film f:

- *volume* (*vol*): number of tweets regarding the film,
- *averaged polarity score* (*aps*): mean of all the raw polarity scores of the tweets concerning the film,
- *positive volume* (*pv*): volume of positive tweets according to a given model for the film,
- *negative volume* (*nv*): $vol - pv$ (our classifiers are binary).

`vol` is a single feature whereas `aps`, `pv` and `nv` are computed for every sentiment models (27 models are used). A total of

82 features is created, all details about the different models are given in the experimental section.

The analytical solution of equation (2) is computed according to: $\mathbf{w}_r^\star = (X_r^T X_r + \lambda I)^{-1} X_r^T Y_r$ with $X_r = \begin{bmatrix} \mathbf{x}_{r,1} \\ \dots \\ \mathbf{x}_{r,32} \end{bmatrix}$. We will show in the experimental part that this problem is particularly ill posed since all variables have high correlations. Indeed, sentiment models are very similar and they generate correlated features. As a consequence, regularization is inevitable, as well as strong variable selection and numerical stabilization in the optimization process. The latter consists in using LU factorization to solve (2) instead of basic matrix inversion. The two first points are discussed in the next section.

3.5 Evaluation, Variable Selection & Regularization

As far as the evaluation is concerned, we use a *leave-one-out* (LOO) procedure. The dataset is too small to build a separated test set and we choose LOO to minimize the over-fitting. All results regarding box office forecasting are computed using LOO. The variable selection process also relies on the LOO criterion to build the most efficient variable subset.

In our optimization problem, the number of features ($d_r = 82$) exceeds the number of instance ($N_f = 32$). And, as mentioned previously, variables are bloc-correlated (the 27 used models give close `aps`, `pv` and `nv` for each film. This is a two-fold issue for our system: we have to get rid of useless information to obtain an optimal forecasting but we are not able to deal with the whole set of variable (it then requires a so strong regularization that the variable contributions are not evaluated reliably).

We propose to use a forward stepwise greedy procedure:

1. All subsets of 1 variables are evaluated (leave-one-out criterion).

2. The most interesting variable is added (definitely) to the active set.

3. The procedure is relaunched and one new variable is added at each step.

At the end of the process, we keep the most efficient subset. Such a procedure is likely to slightly over-fit the data as no virgin test set is kept but we think that it is the most efficient approach to this small problem. We admit that the presented performances are probably slightly over-estimated.

The regularization term in equation (2) answers two problems: firstly, it acts as a stabilizer for the optimization stage. Secondly it guaranties a better generalization and enables us to improve the global leave-one-out performance of the system. The λ tradeoff is optimized by line search.

We offer 2 error metrics to evaluate our performances: we compute the error percentage (which is classical for regression problems). However, the box-office of the different films belong to different scales (ranging from ratios 1 to 1000, details in table 4). As a consequence we suggest a second metric: we define 10 categories (regular partitions ranging from the worst film to the greatest success) and we compute a classification error.

4. EXPERIMENTS : CROSS-MEDIA ADAPTATION

In this section, we focus on cross-media adaptation, *i.e.* adapting model learned on a certain type of labeled data (some reviews) to classify data of an other type (some tweets). We present the datasets that have been used and then we compare different models (SVM, SCL and FEDA) learnt on different training sets.

4.1 Datasets & Models

To set up our cross-media adaptation models, we rely on many available labeled review datasets as well as 2 Twitter manually labeled corpus. They refer to different domains, as described in table 1. Obviously, Twitter sets are much smaller due to the cost of the labeling.

Source Datasets	Size
Amazon DVD [3]	10k
Amazon Books [3]	10k
Amazon Kitchen [3]	10k
Amazon Electronics [3]	10k
ACL IMDb [18]	50k
Trip Advisor [34]	50k
Amazon Huge DVD [13]	450k
Amazon Huge Books [13]	1.9M
Amazon Huge Kitchen [13]	70k
Amazon Huge Electronics [13]	140k
Twitter Sanders [30]	1081
Target Dataset	**Size**
Golden Standard [4]	251

Table 1: Datasets with their associated size.

Then we used three models from the literature to compute the sentiment score of each tweet : We used standard transfer (learning on one or more sources, and testing without adaptation on another) as a baseline, FEDA and SCL (as described in section 3).

4.2 Baseline

We established our baseline learning mono and multi-source models without explicit transfer (namely using basic SVM). Our results are presented in Table 2. We must mention that the Golden Standard is not balanced (82% of positive tweets). To be fair, this prior is given in the learning stage of every models (unbalanced cost option).

The best model is obtained using all the data for learning : we meet the conclusion of [21]. Our best accuracy, 87.25%, is significantly better than the Golden Standard balancing: reviews are efficient for this task. In comparison, the Twitter based model performs poorly, probably due to domain adaptation problem. Sources that are both large and movie related (IMDb and Amazon huge DVD) seem more efficient on the Golden Standard (once more it is probably due to the lack of domain adaptation problem). TripAdvisor does not enable us to get a good accuracy on Twitter but keep in mind that it does not penalize the joint model IMDb+TripAdvisor.

We notice in Table 2 that huge Amazon gives better results than standard Amazon whatever the domain is. In order to study the influence of the learning set size, we consider the global merged training set and we carry out a new series of experiments: we learn some models on a fraction of

Source Dataset	Accuracy on the target
Amazon DVD	82.47%
Amazon Books	82.87%
Amazon Kitchen	82.07%
Amazon Electronics	82.07%
Amazon all	82.87%
ACL IMDb	84.46%
Trip Advisor	75.7%
ACL IMDb & Trip Advisor	84.46 %
Amazon Huge DVD	84.06%
Amazon Huge Books	83.27%
Amazon Huge Kitchen	83.67%
Amazon Huge Electronics	84.46%
Twitter Sanders	64.14%
All	**87.25%**

Table 2: Accuracies on the Golden Standard [4] with respect to the source with basic SVM.

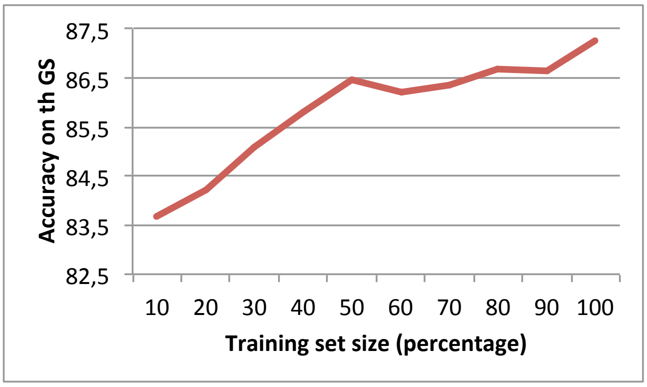

Figure 1: Evolution of the accuracy on the Golden Standard with respect to the percentage of data used for learning (All training set merged). Each point is averaged over 5 experiments.

the whole set and we evaluate the accuracy on the Golden Standard. Figure 1 shows clearly that until a point, the performance increases linearly with the training set size.

4.3 Explicit Adaptation

Our second set of experiments use FEDA and SCL algorithms (cf section 3). We used different sources and the Twitter Sanders dataset [30] as target[3]. Our results are presented in Table 3. The accuracies slightly overcome the previous baselines but the difference is not significant given the small size of the Golden Standard.

Our conclusion is that it is more important to get a large training set than to implement a complex explicit transfer strategy.

5. EXPERIMENTS : BOX OFFICE FORECASTING

In this series of experiments, we focus on the box-office prediction. Previous articles on this subject rely mainly on *expert* features (i.e. film budget, actor celebrity, number of cinemas at the film release...) [6, 31, 33] and Twitter is used only to measure the volume of discussion around the film. We do not used the expert informations and we only focus on Twitter. We want to show the interest of sentiment analysis on Twitter on top of volumetric analysis.

Learning Dataset	FEDA	SCL
Amazon all	81.67%	82.47%
ACL IMDb	82.87%	**86.06%**
Trip Advisor	75.3%	73.71%
Amazon Huge All	**87.85%**	**87.06%**
All	**87.45%**	**87.06%**

Table 3: Accuracies on the Golden Standard [4] using explicit transfer models (FEDA and SCL).

5.1 Dataset & sentiment models

We work on the Twitter dataset from [4]. Once collected[4], the dataset counts 168032 tweets regarding 32 films. The volumetric distribution is given in table 4. The box office is given in dollar, figures come from `www.boxofficemojo.com`.

Title	♯ tweets	Box Office
Edge Of Darkness	3910	43313890
When In Rome	3271	32680633
Tooth Fairy	3111	60022256
Book Of Eli	5845	94835059
Legion	4863	40168080
Extraordinary Measures	799	12068313
Spy Next Door	1934	24307086
To Save A Life	922	3777210
Preacher's Kid	483	515065
Dear John	11229	80014842
From Paris With Love	3137	24077427
Valentine's Day	5335	110485654
Wolfman	3455	61979680
Shutter Island	20229	128012934
Cop Out	4628	44875481
Crazies	2602	39123589
Ghost Writer	2665	15541549
Alice In Wonderland	29112	334191110
Diary Of A Wimpy Kid	1211	64003625
Bounty Hunter	5968	67061228
She's Out Of My League	2474	2010860
Our Family Wedding	1073	20255281
How To Train Your Dragon	5728	217581231
Back Up Plan	955	37490007
Date Night	9041	98711404
Death At A Funeral	3232	42739347
Clash Of The Titans	10547	163214888
Last Song	5702	62950384
Iron Man 2	7075	312433331
My Name Is Khan	4941	4018771
Brooklyn's Finest	2	27163593
Shrek Forever After	2549	238736787

Table 4: Films with their associated number of tweets in [4] and box office in dollars.

Then we generate 82 features using 27 models which are described in table 5. The different datasets used come from classical that are mentioned in the table. Three particularly small experiments will have a great influence on the results: feature 15 is learnt on tweets based on a sentiment criterion, features 16 and 17 focus on another task, discriminating neutral and opinionated documents. [25] contains

[3] Due to the small size of the Golden Standard, it was not possible to split it into a test set and a validation set. As a consequence, we use the closest available set (Sanders Twitter) to perform the explicit adaptation.

[4] Tweets are referenced using the Twitter id and some of them are no longer available.

5000 sentences that have been labeled manually as objective or subjective. For feature 17 we use the neutral labels from the Golden Standard used in the previous section. It should also be noted that features 18 to 22 use the Golden Standard (without labels) and that features 23 to 27 use the Golden Standard with the labels.

ID	Training Source	♯ set	Algorithm
1	IMDB [18]	50k	SVM
2	TripAdvisor [34]	50k	SVM
3	IMDB+TripAdvisor	100k	SVM
4	Amazon (books) [3]	10k	SVM
5	Amazon (dvd) [3]	10k	SVM
6	Amazon (electronics) [3]	10k	SVM
7	Amazon (kitchen) [3]	10k	SVM
8	Amazon (full) [3]	40k	SVM
9	Huge Amazon (books) [13]	1.9M	SVM
10	Huge Amazon (dvd) [13]	450k	SVM
11	Huge Amazon (electronics) [13]	140k	SVM
12	Huge Amazon (kitchen) [13]	70k	SVM
13	Huge Amazon (full) [13]	2.5M	SVM
14	All	2.8M	SVM
15	Twitter Sanders [30]	1081	SVM
16	IMDB Subj/Obj [25]	5000	SVM
17	GS Subj/Obj [4]	754	SVM
18	IMDB [18]	50k	SCL
19	TripAdvisor [34]	50k	SCL
20	Amazon [3]	40k	SCL
21	Huge Amazon [13]	2.5M	SCL
22	All	2.8M	SCL
23	IMDB [18]	50k	FEDA
24	TripAdvisor [34]	50k	FEDA
25	Amazon [3]	40k	FEDA
26	Huge Amazon [13]	2.5M	FEDA
27	All	2.8M	FEDA

Table 5: Models and associated training sets description.

5.2 Baseline & best results

First, we propose to estimate the box office using only the volume of tweets (cf Table 4). Once the volume rescaled (no learning here), we obtain the Figure 2. It represents an averaged error of 210%. According to our second error metric (cf end of section 3.5), this baseline gives a classification error of 56% (over 10 categories of box-office volume).

Then we operate the variable selection process as well as the optimization of the regularization parameter λ and we obtain the following result: 25% of averaged error (for the prediction of the box-office value) and 10% error for the film classification. It is clear that the sentiment features enable us to greatly improve our model. Even if the optimization process generates a slight over-fitting[5], the gap between the two approaches is really significant and demonstrates the usefulness of the sentiment analysis for this task.

5.3 Variable selection & regularization setting

As previously stated (cf section 3.5), variables are highly correlated and it is not possible to efficiently apply a regression technique on raw data. We use a forward stepwise selection procedure which is greedy and based on the leave-one-out evaluation criterion. Figure 3 illustrates the evolution of the error with respect to the number of variables that are used.

Best performance corresponds to 34 variables (25% of averaged error on the box-office value and 10% error for the film classification). The curve is not smooth at all, this is due to the low regularization tradeoff that is used here. Three points should be discussed further:

- using one variable leads to 600% error whereas baseline volumetric model offers a better result (210%). This can be explained easily: baseline model consists in a normalization of the volumetric data which is done over the whole dataset whereas experiments of Figure 3 are based on leave-one-out which is less biased (and harder to optimize).
- studying the models that are selected by the algorithm is interesting: the 10 first variables (corresponding to the first plateau of performance) are the following: 16 5 15 2 17 16 7 3 4 10 (cf Table 5). Objective/Subjective classifiers are used three times, Twitter models is used once and Dvd/IMDB based models are used five times: given the data nature, this result seems logical but do notice that neither the volumetric feature nor the Golden Standard based feature are picked in this first series.
- studying the contribution of each feature reveals some interesting pieces of information: all 34 features are balanced between `aps` (averaged positive scores) and `nv` (negative volume) but no positive volume is used. `aps` and `pv` are probably too close to be used together.

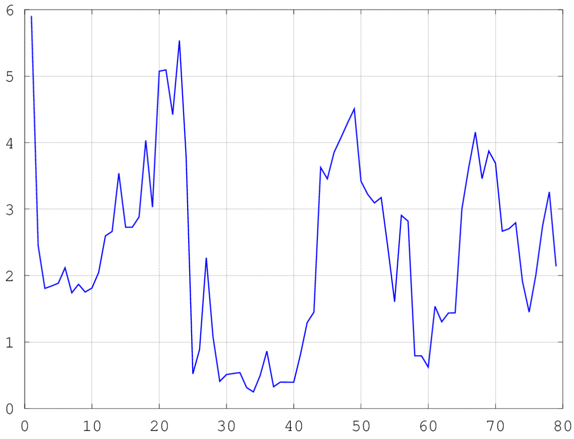

Figure 3: Average error on box-office prediction (in percentage) with respect to the number of features selected.

As far as the regularization setting is concerned, we perform a line search. The result is presented in Figure 4. The best performance (according to both criteria, cf end of section 3.5) corresponds to $\lambda = 1e - 5$.

6. CONCLUSION

In this article we take up on two problems: cross-media sentiment classification and box office forecasting. We draw several conclusions from those 2 series of experiments.

[5] No virgin test set is used, the presented results are obtained using leave-one-out procedure.

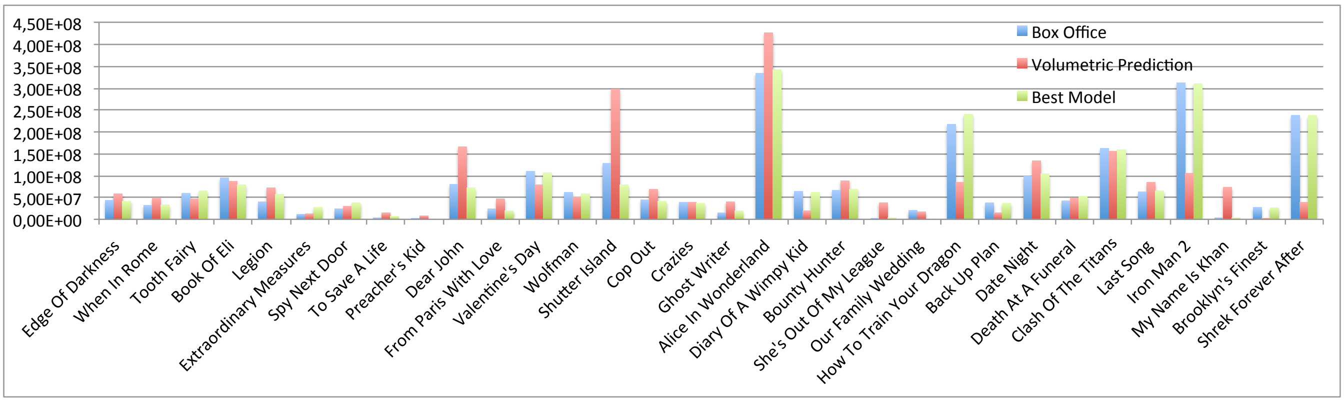

Figure 2: Comparison of the real box office of 32 films with the baseline model (volumetric prediction) and best model (regression over volumetric and sentiment features).

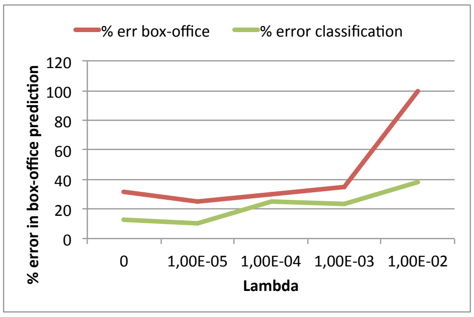

Figure 4: Evolution of the 2 error criteria with respect to the regularization tradeoff λ.

Concerning the cross-media issue, we compare actually two kinds of approaches: first, we propose to use larger training set to face efficiently the unknown Twitter data. Secondly, we test the interest of explicit transfer model on this task. We demonstrate that the size of the learning set is more influent than the dedicated transfer model: this conclusion is interesting since opinionated data resources are almost infinite with the web 2.0.

Regarding box-office forecasting, we show the interest of sentiment features to improve the performance. We also study the interest of the different variables and we conclude that we need features from different problems: Objective/Subjective discrimination, Positive/Negative classification, ... Each feature is linked to a model: the variable selection process do not focus on best sentiment models (according to the first task) but it selects models related to movies and tweets as well as Objective/Subjective discriminators.

Finally, the two considered tasks are less linked than expected: the cross-media sentiment classification requires large training sets whereas box-office prediction requires a large range of opinion mining derived features.

7. ACKNOWLEDGEMENTS

The research reported herein was supported by the DIFAC project and the AMICO project.

8. REFERENCES

[1] S. Asur and B. A. Huberman. Predicting the future with social media. In *ACM IC Web Intelligence*, 2010.

[2] D. Bespalov, B. Bai, Y. Qi, and A. Shokoufandeh. Sentiment classification based on supervised latent n-gram analysis. In *ACM CIKM*, pages 375–382, 2011.

[3] J. Blitzer, M. Dredze, and F. Pereira. Biographies, Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *ACL*, 2007.

[4] L. Chen, W. Wenbo, M. Nagarajan, S. Wang, and A. P. Sheth. Extracting diverse sentiment expressions with target-dependent polarity from twitter. In *ICWSM*. The AAAI Press, 2012.

[5] H. Daumé-III. Frustratingly easy domain adaptation. In *ACL*, 2007.

[6] C. Dellarocas, X. M. Zhang, and N. F. Awad. Exploring the value of online product reviews in forecasting sales: The case of motion pictures. *Journal of Interactive Marketing*, 21(4):23–45, 2007.

[7] X. Ding, B. Liu, and P. S. Yu. A holistic lexicon-based approach to opinion mining. In *ACM WSDM*, pages 231–240, 2008.

[8] M. Dredze, A. Kulesza, and K. Crammer. Multi-domain learning by confidence-weighted parameter combination. *Machine Learning Jour.*, 79(1–2):123–149, 2010.

[9] M. Dredze, A. Kulesza, and K. Crammer. Multi-domain learning by confidence-weighted parameter combination. *Machine Learning*, 79:123–149, 2010.

[10] S. Gerrish and D. Blei. Predicting legislative roll calls from text. In *ICML*, pages 489–496, 2011.

[11] X. Glorot, A. Bordes, and Y. Bengio. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *ICML*, 2011.

[12] M. Hu and B. Liu. Mining and summarizing customer reviews. In *ACM SIGKDD*, pages 168–177, 2004.

[13] N. Jindal and B. Liu. Opinion spam and analysis. In *ACM WSDM*, 2008.

[14] T. Joachims. *Learning to Classify Text using Support Vector Machines*. Springer - Kluwer Academic Publishers, 2002.

[15] M. Joshi, D. Das, K. Gimpel, and N. A. Smith. Movie reviews and revenues: An experiment in text regression. In *In Proceedings of NAACL-HLT*, 2010.

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

[17] B. Liu. Sentiment analysis and subjectivity. In *Handbook of Natural Language Processing, Second Edition. Taylor and Francis Group, Boca*, 2010.

[18] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts. Learning word vectors for sentiment analysis. In *Association for Computational Linguistics (ACL)*, 2011.

[19] Y. Mansour, M. Mohri, and A. Rostamizadeh. Domain adaptation with multiple sources. In *NIPS*, 2008.

[20] J. Marcoux and S.-A. Selouani. A hybrid subspace-connectionist data mining approach for sales forecasting in the video game industry. *Computer Science and Information Engineering*, 5:666–670, 2009.

[21] Y. Mejova and P. Srinivasan. Crossing media streams with sentiment: Domain adaptation in blogs, reviews and twitter. In *ICWSM'12*, 2012.

[22] P. Melville, W. Gryc, and R. D. Lawrence. Sentiment analysis of blogs by combining lexical knowledge with text classification. In *Proceedings of the 15th ACM SIGKDD*, pages 1275–1284. ACM, 2009.

[23] A. Pak and P. Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, 2010.

[24] S. Pan, X. Ni, J.-T. Sun, Q. Yang, and Z. Chen. Cross-domain sentiment classification via spectral feature alignment. In *WWW*, 2010.

[25] B. Pang and L. Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *ACL*, 2004.

[26] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Information Retrieval*, 2:1–135, 2008.

[27] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *ACL-Empirical Methods in NLP*, volume 10, pages 79–86, 2002.

[28] V. M. K. Peddinti and P. Chintalapoodi. Domain adaptation in sentiment analysis of twitter. In *Analyzing Microtext*, AAAI Workshops. AAAI, 2011.

[29] A. Rafrafi, V. Guigue, and P. Gallinari. Coping with the document frequency bias in sentiment classification. In *AAAI ICWSM*, 2012.

[30] N. J. Sanders. Twitter sentiment corpus, 2011.

[31] I. Shtrimberg. Good news or bad news? let the market decide. In *In AAAI Spring Symposium on Exploring Attitude and Affect in Text. Palo Alto: AAAI*, pages 86–88. Press, 2004.

[32] S. Tan, X. Cheng, Y. Wang, and H. Xu. Adapting naive bayes to domain adaptation for sentiment analysis. In *ECIR*, volume 5478, pages 337–349. Springer, 2009.

[33] A. Tumasjan, T. Sprenger, P. Sandner, and I. Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, pages 178–185, 2010.

[34] H. Wang, Y. Lu, and C. Zhai. Latent aspect rating analysis on review text data: A rating regression approach. In *ACM SIGKDD*, pages 783–792, 2010.

[35] M. Whitehead and L. Yaeger. Building a general purpose cross-domain sentiment mining model. In *IEEE Computer Science and Information Engineering*, pages 472–476, 2009.