Crossing Media Streams with Sentiment: Domain Adaptation in Blogs, Reviews and Twitter

Yelena Mejova and **Padmini Srinivasan**
Computer Science, University of Iowa, Iowa City, IA
yelena-mejova, padmini-srinivasan@uiowa.edu

Abstract

Most sentiment analysis studies address classification of a single source of data such as reviews or blog posts. However, the multitude of social media sources available for text analysis lends itself naturally to domain adaptation. In this study, we create a dataset spanning three social media sources – blogs, reviews, and Twitter – and a set of 37 common topics. We first examine sentiments expressed in these three sources while controlling for the change in topic. Then using this multi-dimensional data we show that when classifying documents in one source (a *target* source), models trained on other sources of data can be as good as or even better than those trained on the target data. That is, we show that models trained on some social media sources are *generalizable* to others. All source adaptation models we implement show reviews and Twitter to be the best sources of training data. It is especially useful to know that models trained on Twitter data are generalizable, since, unlike reviews, Twitter is more topically diverse.

Introduction

When we speak of *Social Media* today, we refer to a wide variety of social forums: blogs, wikis, review forums, social networking sites, and many others. Social media are of interest to text mining researchers, especially since these present a glimpse into the thoughts, viewpoints and feelings of a vast number and variety of people. A branch of study that has attracted a lot of attention is the identification of sentiment in social media such as in user blog posts, tweets and reviews.

However, despite the progress made, there are key areas in sentiment analysis (SA) research that have not received sufficient attention. First we observe that a vast majority of SA studies center around a single source of data, such as blogs, or web pages, or reviews. There is a related strand of research in topic discovery and modeling, such as in (Paul and Girju 2009) comparing topics extracted from different social media streams. But these do not focus on sentiment analysis. This is only changing recently with researchers addressing SA questions that involve multiple sources of data. For instance, (Bermingham and Smeaton 2010) asks whether it is easier to classify sentiment of short documents like Twitter than longer ones like blogs. In these few studies, though, the data from various sources are sampled and collected using different strategies, thus permitting very limited comparisons. For example, we still do not know if different streams project different sentiment expressions on the same topic.

Which leads to our second key observation that studies involving multiple sources (we use stream and source synonymously) make little or no attempt to control for topic. So, given a particular topic (e.g., *Motorola Droid*) we do not know if a) the different sources discuss the topic more or less to the same extent, and furthermore b) if the sentiments expressed on the topic are similar or not. A question to ask here, for instance, would be – do Twitter users react the same way to a specific news item as, say, bloggers or Facebook users? This type of question has not yet been posed. One could make the assumption that these reactions are likely to be similar and proceed forward. However, these social media streams differ categorically in the kinds of input and interactions supported. A blogging medium allows long, thoughtful posts (amongst other varieties of course), Twitter with its size constraints vastly differs in this regard. YouTube is video driven while still often invoking strong sentimental reactions as seen in comments. Facebook, offers a different mode of interaction with its feeds, etc. These distinctions may be sufficient to create significant differences in the core populations they attract. Hence the assumption about similarities in reaction to a topic may quite easily be challenged.

We seek to understand this aspect by comparing sentiment expressions across sources *while controlling for topic*. A better understanding will indicate whether it is reasonable to rely on a single source when the goal is to mine opinion on a topic. Thus in this paper we compare sentiment over three particular social media streams: blogs, microblogs (Twitter) and reviews. We show, for example, that roughly 50 to 70% of blogs, depending upon topic category, convey positive sentiment while with Twitter this drops considerably to roughly 30 to 50%. And there are category-specific polarity differences: compare 54% positive documents found in blogs to 27% in Twitter for computer games, 47% positives in blogs and 30% in Twitter for movies.

A second major goal in this paper is to explore how best to build SA classifiers when we have data along a two-

Copyright © 2012, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

dimensional grid, one varying over source and the other on topic. In particular we use this data to explore stream (or source) adaptation and answer the question: *to what extent can a SA classifier developed on a topic for one social medium be transferred for use on a different medium?* There is related prior work, as for example, in (Blitzer, Dredze, and Pereira 2007), but instead of being cross source, they explore cross topic SA classifiers within the same medium. For example they train classifiers on documents about electronics and test on other documents about movies. In contrast, (Peddinti and Chintalapoodi 2011) do explore cross-source adaptation, but only from microblogs to reviews, and again without topical constraints on the datasets. The difference in our work is that we study cross-source (aka domain) classifiers for the same topic. We explore both single-source and multiple-source adaptation. For example, we study SA classifiers trained on reviews and blogs and tested on tweets for the same topic. We also study voting approaches as another angle for combining SA classification knowledge. We show, for example, that Twitter, despite its size restrictions, is a good source for building classifiers to be used in other kinds of data.

The specific contributions of our work are:

1. We analyze sentiment across media for various topics and topic categories. We consider positive, negative, neutral, and mixed classes of sentiment.
2. We conduct a set of cross-source adaptation experiments for building sentiment classifiers using
   (a) single-source models,
   (b) multiple-source mixed models, and
   (c) multiple-source voting models.
3. We contribute a novel multi-source, multi-topic labeled dataset geared specifically for comparing sentiment across social media sources.

In the following sections we provide a brief overview of work related to sentiment analysis and source domain adaptation, describe how our dataset was created, including the annotation strategy. We then analyze our data streams for topicality and sentiment. Finally, we present our domain adaptation experiments and discuss the implications of their results.

Related Work

Much of early SA centered around product reviews, such as ones left for products on Amazon.com. This was convenient since the star ratings provide a quantitative label for the sentiment expressed in the documents (making it unnecessary to manually label them) (Pang and Lee 2002). Later, more general types of writing such as blogs (Melville, Gryc, and Lawrence 2009), web pages (Kaji and Kitsuregawa 2007) and news articles (Tan et al. 2009) were explored. Recent growth of the micro-blogging site Twitter has also produced a plethora of research tracking topics and their sentiment (Asur and Huberman 2010; Chew and Eysenbach 2011). Cross-topic adaptation in the context of sentiment analysis, but within the same source, has been widely studied. For example, (Blitzer, Dredze, and Pereira 2007) use a dataset of four topical categories of Amazon product reviews to gauge the "similarity" between the topics and their potential for classifier adaptation. The same dataset is used by (Mansour, Mohri, and Rostamizadeh 2008) to test their theoretical analysis of adaptation. More recently, (Tan et al. 2009) and (Pan et al. 2010) have looked at the topic-level adaptation of sentiment classifiers but, again, only in the context of reviews. Other social media sources have also been studied in this fashion. For instance, (Chesley et al. 2006) mine news, blogs, and even letters to newspaper editors.

More generally, (Daume 2007) explores the use of a variety of sources, including blogs, news, usenet, and conversational telephone speech for sequence labeling tasks such as named-entity recognition, shallow parsing, and part-of-speech recognition. Cross-language sentiment classification has also been performed by (Wan 2009) on Chinese and English reviews.

Few studies have compared sentiment expression and classification in different social media sources. (Bermingham and Smeaton 2010) examine classification performance on review and blog sources in which documents are constrained in length, finding that it is easier to classify shorter documents than their longer counterparts. Again they do not fix the topic across streams. Closer to our work, (Peddinti and Chintalapoodi 2011) develop iterative algorithms for filtering noisy data during source adaptation from Twitter and Blippr (a micro-review) to movie reviews. Unlike our study, they do not control for topical coverage during data selection, nor do they examine the sentiment expressed in their data.

Thus, to the best authors' knowledge at the time of the writing, our approach to gauging cross-source classifiers across blogs, reviews, and Twitter using a multi-dimensional dataset spanning more than 30 topics in five categories is a new contribution to the field.

Data Streams

We explore sentiment across three social media sources – blogs, microblogs (Twitter), and reviews. Reviews tend to be the most topic focused. Tweets also tend to be topic focused within their size restrictions. Individual blog posts may contain any number of topics, sometimes unrelated to each other. Thus, each "stream" provides a different outlet for self-expression and topic discussion, potentially affecting the sentiment expressed in each. Note, however, that our work is different from research on synchronized streams popular in the emerging topic detection or topic tracking literature, since we do not control for when the documents were published during collection.

Data Collection

Our data collection procedure aims for roughly equal topic representations, i.e., retrieved sets in all three streams for each of our five topic categories. The categories are *movies*, *music albums*, *smart phones*, *computer games*, and *restaurants*. Each category consists of several topics gathered from outside authoritative sources and pruned during data collection. The data was then cleaned and sampled. The blog

and Twitter subsets were iteratively labeled for topic relevance and sentiment using Amazon Mechanical Turk. The resulting datasets provide us with the texts to analyze for the common set of topics across three social media sources.

An initial set of topics were identified, for movies from Internet Movie Data Base (imdb.com), musical albums from Amazon (amazon.com), computer games from Metacritic (metacritic.com), phones from CNet (cnet.com), and restaurants from Yelp (yelp.com). Starting at the top of each initial list of topics (ranked by popularity), we retrieve and collect documents using the following 2-part rule:

- if # of returned results from any stream is $< 50 \rightarrow$ discard the topic, else keep the topic
- if # of returned results from a stream is $> 100 \rightarrow$ select 100 randomly

The topics passing the above rules are retained in their topical category. We iterate through topics until we have retrieved a minimum of 500 documents in each stream per topical category. The final collection is then cleaned using stream-specific approaches as described below. These parameter values were selected so as to keep the datasets both meaningful in supporting our goals and also manageable for sentiment annotation. This strategy produced a dataset which is in size comparable to ones in (Blitzer, Dredze, and Pereira 2007; Bermingham and Smeaton 2010).

To collect Twitter messages, we used Search API, excluding tweets of less than 10 characters in length. For blogs we used Google Blog Search API and crawled the pages, cleaning HTML and extracting the content using heuristics. To extract post content we first look for the title (given by Google API), and analyze the text after it to get the content in which there are relatively few HTML tags. Specifically, as we process the text we keep track of a *tag density* measure, conveying to us how much text is shown compared to the number of HTML tags. We start collecting blog post content when there are five consecutive words without HTML tags and stop when tag density spikes. The gathered text also must consist of at least 90% alphanumeric characters, and must be at least 100 characters in length. These parameters were selected empirically. We collected reviews by scraping various websites, coding specific scrapers for IMDB.com, CNet.com, yelp.com and Amazon.com[1]. Except for twitter which constrains retrieval to the past 2 weeks, there was no attempt to constrain the other two streams to a particular time period.

Data Labeling

We use Amazon Mechanical Turk[2] (AMT) to label topic relevance and sentiment in the blog and Twitter subsets. Providing a marketplace for work that requires human intelligence, such as data labeling, AMT has become popular in information retrieval and machine learning research (Sheng, Provost, and Ipeirotis 2008). However, rife with bots and some users not providing quality work, data gathered on AMT must be cleaned and quality control set in place.

[1]Reviews were randomly sampled.

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

Aiming to collect three ratings for each document, we designed two tasks (or Human Intelligence Tasks – HITs), one for blogs and another for tweets. Only raters with approval ratings over 90% were allowed to participate. Each blog HIT contained 5 blog posts and Twitter HIT 10 tweets. At the end of each task the annotator is asked to enter the first word of the last document in the HIT as a quality control measure. Within the 10 tweets we also insert a "control" tweet with an obvious sentiment polarity. If either control is failed, the whole HIT is rejected. The tasks were published in stages. HITs rejected by our quality controls during the first stage of rating were re-published. Only two such stages were necessary to collect 99% of the desired HITs.

Raters were asked to annotate each document (blog post or tweet) first for topicality – whether the document is relevant to the topic – with available choices being *Yes*, *No*, and *Can't Tell*. For relevant documents the raters were asked to identify the sentiment it expresses toward the topic: *Positive*, *Negative*, *Mixed*, *None*, or *Can't Tell*.

We calculate inter-annotator agreement using a technique designed specifically for AMT tasks (Snow et al. 2008). This special measure must be used because several hundred users took part in the labeling process, and standard measures such as Cohen's kappa would not be applicable to this set up. Defined in a prominent study of Amazon Mechanical Turk (Snow et al. 2008), this measure is calculated by averaging Pearson correlation for each set of ratings with the average rating. We analyzed the labeling process in three stages. First, annotators had to decide whether the document was on topic. The agreement on this task was 0.600 for blogs and 0.389 for Twitter. This suggests the need for more precise retrieval strategies for Twitter. Next, the topical documents had to be rated according to their sentiment. The agreement on whether the document had sentiment (i.e. was subjective) was at 0.260 for blogs and 0.490 for Twitter. Finally, the task of distinguishing positive from negative documents had an agreement of 0.305 for blogs and 0.535 for Twitter. Not surprising blogs proved to be more challenging in sentiment classification task than Twitter. A subset – 10 Twitter and 10 Blog HITs – were rated by an expert not associated with the project, and ratings compared to the majority rating. A similar difficulty level was seen with 67.7% of Twitter and 58.0% of blog annotation overlap.

Stream Characteristics: Topicality and Sentiment

We now address the question: *What are the differences in sentiment expressed in blog, Twitter, and review streams?* Table 1 shows the topicality and sentiment characteristics of the three streams. The final labels were decided using majority vote of the three ratings each document has received. Documents with no clear majority appear under Other. The column also includes entries marked *Can't tell*. The division between Pos, Neg, and Mix classes for reviews was done according to the star ratings. For five-star ratings we took 1-2 as Neg, 3 as Mix, and 4-5 as Pos. For ten-star ratings we took 1-3 as Neg, 4-7 as Mix and 8-10 as Pos. The percentages (in parentheses) for the topical classes are those of the

Table 1: Distribution of topical and sentiment documents. Percentages are in parentheses.

	Blogs								
Category	Total	Topical	Not top.	Other	Pos	Neg	Mix	None	Oth
Movies	423	184 (44)	196 (46)	43 (10)	87 (47)	12 (7)	29 (16)	46 (25)	10 (5)
Music	462	243 (53)	160 (35)	59 (12)	154 (63)	8 (3)	19 (8)	51 (21)	11 (5)
Games	525	285 (54)	187 (36)	53 (10)	154 (54)	20 (7)	32 (11)	60 (21)	19 (7)
Phones	427	261 (61)	136 (32)	30 (7)	130 (50)	17 (7)	33 (13)	60 (23)	21 (8)
Rest-nts	355	138 (39)	172 (49)	45 (12)	96 (70)	2 (1)	17 (12)	15 (11)	8 (6)
Total	2192	1111 (51)	851 (39)	230 (10)	621 (56)	59 (5)	130 (12)	232 (21)	69 (6)
	Twitter								
Category	Total	Topical	Not top.	Other	Pos	Neg	Mix	None	Oth
Movies	770	612 (80)	126 (16)	32 (4)	182 (30)	41 (7)	16 (3)	319 (52)	54 (9)
Music	740	731 (99)	3 (0)	6 (1)	263 (36)	10 (1)	10 (1)	397 (54)	51 (7)
Games	495	473 (95)	14 (3)	8 (2)	128 (27)	26 (6)	42 (9)	231 (49)	46 (10)
Phones	482	479 (99)	1 (0)	2 (1)	187 (39)	99 (21)	29 (6)	142 (30)	22 (5)
Rest-nts	566	545 (96)	9 (2)	12 (2)	268 (49)	14 (3)	32 (6)	200 (37)	31 (6)
Total	3053	2840 (93)	153 (5)	60 (2)	1028 (36)	190 (7)	129 (5)	1289 (45)	204 (7)
	Reviews								
Category	Total	Topical	Not top.	Other	Pos	Neg	Mix	None	Oth
Movies	800	800 (100)	–	–	612 (77)	91 (11)	97 (12)	–	–
Music	772	772 (100)	–	–	627 (81)	84 (11)	61 (8)	–	–
Games	617	617 (100)	–	–	504 (82)	63 (10)	50 (8)	–	–
Phones	500	500 (100)	–	–	316 (63)	96 (19)	88 (18)	–	–
Rest-nts	900	900 (100)	–	–	715 (78)	70 (8)	115 (13)	–	–
Total	3589	3589 (100)	–	–	2774 (77)	404 (11)	411 (12)	–	–

total, and for sentiment classes are of the total number of topical documents.

Notice that Twitter generally returned larger numbers of documents of which a minimum of 80% were marked topical. For blogs on the other hand only 39 to 54% of the retrieved documents were topical. Intuitively, it makes sense that the longer documents such as blog posts would have more noise which would disrupt information retrieval. Our data distribution underlines the difficulty of retrieving topical blogs and the comparative ease of retrieving from Twitter. In terms of raw numbers too, Twitter appears as a rich stream, supporting its recent wide use for topic tracking (Kouloumpis, T., and Moore 2011; Davidov, O., and Rappoport 2010).

Examining the topical documents we note that positive is the dominant class in all three streams. This trend is consistent across the three streams. In other words, when the effort is taken to convey an opinion on social media on these topics, it is predominantly to convey positive rather than negative impressions. This trend comes across as strong when we limit ourselves to the documents that are judged Pos or Neg. Note that the Mixed class rightfully belongs to both Pos and Neg and so does not impact the relative proportion. We find that reviews are 87% positive, blogs 91% positive and tweets 92% positive. These high percentages also indicate that if we are looking to mine Neg opinions then reviews have some advantage. The other implication of this result is that it is not realistic to build sentiment classifiers from datasets where the Pos and Neg classes are balanced in number (e.g., (Pang and Lee 2002; Blitzer, Dredze, and Pereira 2007)). Our data strongly indicates that there is a significant skew towards the Pos class for these topics and therefore identifying Neg documents is the hard problem.

Additional interesting observations can be made. In Twitter the percentage of topical documents with no sentiment (category None) is typically double that of Blogs (except in Restaurants). This indicates that Twitter is not only a way for people to express their opinions, but is also a way to disseminate purely informational content. Understandably, a non-trivial portion of blogs is of mixed sentiment, making this stream more challenging for sentiment analysis. Reviews too, are not severely constrained in size, and thus allow for a more complex sentiment expressions with 12 percent mixed documents.

There are also a few topic-specific differences. Phones show a large negative presence in tweets and reviews, but not so in blogs. It seems dissatisfied electronics consumers prefer Twitter to express their dissatisfaction. On the other hand, as a share of all topical documents, blogs provide a more complex discussion of movies, having the share of Mix class (16%) more than twice as large as Neg, whereas Twitter's share of mixed documents on the topic is much smaller (3%). Thus, sentiment extracted from each stream must be examined in the light of the stream's general tendencies about particular topics.

We conclude that blogs, reviews, and Twitter each have their own peculiarities when we examine the sentiment of their contents. Of all sentiment-laden documents in the stream, blogs provide the most mixed and fewest negative documents with a proportion of roughly 12 positive to every negative document, whereas Twitter presents a slightly more balanced proportion of roughly five positive to every negative document. Topical differences between the streams

also point to the dangers of using only one stream for gauging public sentiment about a topic. For example, a consumer would find more negative sentiments about phones in tweets than in both blogs or reviews. Thus, within the limits of this dataset, we conclude that blogs, reviews, and Twitter differ significantly as sources of sentiment-laden documents, and may bias studies using only one source.

Cross-Stream Classifier Experiments

Using the above dataset next we consider the question: *how well do sentiment classifiers trained on one stream perform while classifying data from another stream?* Note that as part of our response to this question we also test performance where training and testing are done within a stream.

We use Lingpipe language model-based logistic regression classifier, which has been widely used for sentiment classification (Denecke 2008; Ye, Zhang, and Law 2009). Lingpipe classifier uses its language analysis framework to parse the text and extract n-gram features with $n \leq 3$.

Single-source Model Adaptation

To test cross-stream performance of our classifiers, we perform an evaluation using 3-fold cross-validation (choosing 3 folds instead of more in the light of a small minority class). For each topic/stream combination, we train classifiers on two thirds of the data from one stream, and test on a third of the target stream. We repeat this three times.

We build two binary sentiment classifiers for each topic – one to identify positive documents in the topical set and the other to identify negative documents in the same set. We build 2 classifiers instead of the usual one (classifying positive versus negative) because we have a mixed class of documents. These documents rightfully belong both to positive and negative classes. Therefore we cannot use a single binary classifier. Our design reflects the real-world situation where some documents may contain just positive, just negative, mixed, or even no sentiment at all. Compared to a binary positive/negative classification task popular in the sentiment analysis literature today (Davidov, O., and Rappoport 2010; Pang and Lee 2002), this task is more difficult, but makes fewer assumptions about the nature of sentiment expressions.

Our results are in Table 2. In each cell a classifier was trained on the data specified in the column and tested on data of the row ("target") stream. When the source for building the classifier differs from the target stream we refer to the classifier as 'foreign' and otherwise we refer to it as 'native' (the native runs are underlined). We present both Accuracy and target class F-score as measures. The best performance amongst the three streams for a given target-category-measure combination is in bold. When considering accuracy we see that the best performance within a category - stream combination is mostly achieved by a native classifier. Specifically out of 30 accuracy measurements with native classifiers (5 topical categories * 3 streams * 2 classifiers) 26 native classifiers achieved the highest score. With Target F-score (negative class F-score for negative classifier, positive otherwise), which is the more challenging measure, fewer, i.e., only 19 native classifiers achieved the highest score. Overall these results with native classifiers are not surprising.

Table 2: Single-Source Model Adaptation

			Accuracy			Target F-score		
Task	**Categ.**	**Target**	**Blogs**	**Reviews**	**Twitter**	**Blogs**	**Reviews**	**Twitter**
POS	Games	Blogs	0.631	**0.652**†	0.543	0.763	**0.824**†	0.627
		Reviews	0.788	**0.897**	0.865	0.880	**0.948**	0.927
		Twitter	0.528	0.365	**0.709**	0.520*	**0.646***	0.608*
	Movies	Blogs	**0.655**	0.638*	0.516*	0.783	**0.790***	0.568
		Reviews	0.731	**0.883**	0.759*	0.844	**0.941**	0.861*
		Twitter	0.534	0.367	**0.705**	0.439	**0.612***	0.581
	Music	Blogs	**0.762**	0.708	0.754*	**0.857**	0.848*	0.844*
		Reviews	0.856*	**0.900**	0.878	0.923*	**0.949**	0.937
		Twitter	0.670	0.380	**0.787**	0.558*	0.668*	**0.714**
	Phones	Blogs	**0.635**	0.608*	0.414	0.748	**0.792***	0.410
		Reviews	0.769	**0.815**	0.279	0.869	**0.905**	0.465
		Twitter	0.570*	0.513	**0.610**	0.531*	**0.637**†	0.561
	Rest-nts	Blogs	0.800	**0.814***	**0.814***	0.893	**0.905***	0.896*
		Reviews	0.882*	**0.923**	0.917*	0.938*	**0.961**	0.958*
		Twitter	0.548	0.539	**0.696**	0.651*	**0.753***	0.741
NEG	Games	Blogs	**0.815**	0.794*	0.805*	**0.302**	0.140*	0.126*
		Reviews	0.783	**0.814**	0.809*	0.103	**0.275**	0.037
		Twitter	0.721	0.838*	**0.859**	0.181	0.119	**0.383**
	Movies	Blogs	0.777	0.672	**0.783***	0.096	**0.240**†	0.179*
		Reviews	0.736	**0.802**	0.761*	0.192*	**0.483**	0.236*
		Twitter	0.844	0.800	**0.905**	0.282*	0.185	**0.356**
	Music	Blogs	**0.888**	0.810	0.884*	**0.185**	0.088*	0.000*
		Reviews	0.796	**0.828**	0.811	0.148	**0.435**	0.000
		Twitter	0.953	0.775	**0.978**	0.000*	0.152*	**0.333**
	Phones	Blogs	0.775	0.689	**0.814**†	0.038	**0.250***	0.197*
		Reviews	0.616	**0.698**	0.622	0.141	**0.513**	0.294
		Twitter	0.637*	0.578	**0.704**	0.226*	**0.440***	0.288
	Rest-nts	Blogs	**0.866**	0.837*	0.859*	**0.222**	0.000*	0.000*
		Reviews	0.764*	**0.802**	0.797*	0.186*	**0.354**	0.116*
		Twitter	0.863	0.874*	**0.920**	0.000	0.129*	**0.418**

In contrast, the results look remarkably more interesting when we test differences in performance for statistical significance. We notice that there are many instances in which the performance of a foreign classifier is *statistically indistinguishable* from that of the native classifier – these instances are marked with a *[3]. For example, POS *restaurant* classifiers trained on reviews or on blog posts perform the same (in terms of Accuracy) as the corresponding native classifier. In some cases, as for example the POS *games* classifier trained on reviews even outperforms the native blog-based classifier at a statistical significance of $p < 0.01$! The four classifiers that significantly outperform their native counterparts are marked with †. The number of foreign classifiers which achieve performance statistically indistinguishable from or better than the native classifier (in 43 experiments out of 60) shows that cross-stream adaptation is possible, and in a few cases even beneficial. Thus the answer to our question is that we can, in general, build classifiers on one stream and use it on another. This facility is useful when

[3]In contrast to usual practice given our interest we mark the statistically indistinguishable results.

Table 3: Single-source model adaptation: number of best or statistically indistinguishable from best runs.

		Accuracy		F-score		Either
Source	**Target**	**NEG**	**POS**	**NEG**	**POS**	**All**
Blogs	Blogs	4	3	3	1	7
	Reviews	1	2	2	2	4
	Twitter	1	1	3	4	7
Reviews	Blogs	2	4	5	5	10
	Reviews	5	5	5	5	10
	Twitter	2	0	3	5	9
Twitter	Blogs	5	3	5	2	8
	Reviews	3	2	2	2	5
	Twitter	5	5	4	3	10
	Best possible	5	5	5	5	10

Table 4: Multiple-source model adaptation: number of runs statistically indistinguishable from native classifier.

		Accuracy		F-score		Either
Source Model	**Target**	**NEG**	**POS**	**NEG**	**POS**	**All**
Mixed R + T	Blogs	3	5	5	5	10
Mixed B + T	Reviews	1	0	2	0	2
Mixed B + R	Twitter	3	2	4	4	6
Mixed all	Blogs	5	5	5	4	10
Mixed all	Reviews	5	4	5	5	10
Mixed all	Twitter	5	3	5	1	8
Voting all	Blogs	4	5	5	4	10
Voting all	Reviews	5	3	4	4	9
Voting all	Twitter	5	5	2	5	10
	Best possible	5	5	5	5	10

it is hard to obtain sufficient topical documents in a stream or it is challenging to label documents of a stream. We know from the previous section that blogs were more challenging to label than tweets both in terms of whether they carried sentiment and whether the sentiment was positive or negative. Our cross-stream results indicate that we could use data from other streams to classify blogs.

Examining Table 2 further we observe that if we total the number of times a stream offers the best score or a score that is statistically indistinguishable from the best (in accuracy or the target F-score) then we have the distribution as shown in Table 3. For example, the Blog stream positive classifier offers the best or similar to the best Accuracy in 6 out of 15 experiments (3 target streams * 5 topics). Of these 6, in 3 instances the classifier is a foreign classifier (i.e., classifying reviews or tweets).

It is not surprising to note that reviews are the best stream achieving a total of 29 instances across classifiers and measures with the best or close enough to best performance. Of these, 19 instances are in the role of foreign classifiers. What is most surprising is that Twitter is also a good source of training data, with best or close to best performance in 23 instances and these include 13 instances where the classifier is a foreign classifier. Blogs, on the other hand, offer the best or close enough scores in only 18 instances of which only 11 are as foreign classifiers. Moreover, when blogs or twitter posts are being classified, the native blog classifier is matched by a foreign classifier in 100% of the instances (compared to 60% for reviews).

Thus we infer, within the limits of these experiments, that blogs offer the least interesting classifiers for sentiment, whereas review-based classifiers are the best. Review classifiers offer the best or close enough scores in 10 out of 10 blog classification instances and 9 out of 10 Twitter ones, suggesting that tweets may be slightly more difficult to classify than blog posts. Surprisingly, this is followed by classifiers built on Twitter – a medium that is by design highly constrained in size. To further illustrate Twitter's strength, it offers the best or close enough classifier 5 times out of 10 even while classifying reviews and 8 out of 10 while classifying blogs.

Multiple-Stream Model Adaptation

We further explore cross-stream adaptation by taking advantage of several streams when building a classifier. The question we address is, *does training on several social media sources improve classification performance when adapting to another source?* We explore three scenarios:

- **Two-Source Mixed Model** – a classifier which has been trained on documents from two different streams (excluding target stream)
- **Three-Source Mixed Model** – a classifier which has been trained on three streams (including target stream)
- **Three-Source Voting Model** – three classifiers, each trained on one stream, using majority voting to determine the class of a document

An advantage of using several sources to build sentiment classifiers is the diversity of language and expression the training data includes, compared to training on only one source. We evaluate the performance of these classifiers as in the single-stream experiments above, using 3-fold cross-validation, and compare them to their native counterparts. For each of the above scenaries, ten experiments were conducted: one for each of the five topical categories, times two classifiers (positive and negative).

Figure 1 shows the accuracy of native, 2-source and 3-source mixed, and voting models with 99% confidence intervals. The results are presented first sorted by task (negative and positive), target stream, and finally by topical category. For example, the first interval shows the NEG classifiers used to classify documents on games in the Blog stream. We see that in some instances models match each other very well, such as when detecting negative blog posts (leftmost five tasks). Performances are not as much matched when classifying Twitter, though, especially with 2-source mixed model lagging behind the others. Notably, out of all of these experiments, only in one instance do we get a performance that is statistically better than that of the native classifier – the negative voting model tested on blog posts about phones. Otherwise, the performance is as good as or inferior to the native classifier.

Furthermore, we examine the number of best runs for each model in Table 4. We see that reviews benefit the least

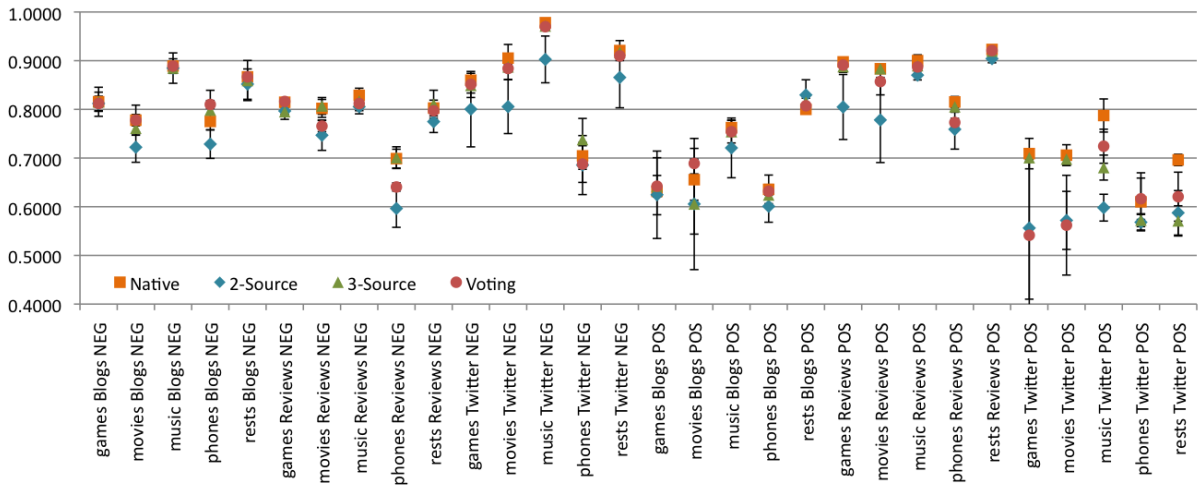

Figure 1: Accuracy of native, 2- and 3-source mixed, and voting models with 99% confidence intervals.

from a 2-mixed source model, followed by Twitter. Once again, blogs are shown to be easiest to classify using foreign training data with a matched performance in 10 out of 10 experiments for all models. Models which used all sources (mixed and voting) perform better than those which exclude the target data. This supports the common intuition that it is always beneficial to train on labeled data for the target dataset whenever possible. Looking closer at the distinction between the voting and mixed models when training on all three streams, we find that the mixed model predicts document class correctly 79.82% of the time compared to 78.57% for the voting model, making the mixed model marginally better.

We conclude that compared to single-source adaptation, it is indeed better to train on many data sources as possible, that is, training on several different sources makes models more comparable to the native model. However, these models may only be comparable to the native model but they will not outperform it. Furthermore, mixing outside data with target data for training classifiers may produce weaker classifiers than if only the target data was used.

Topic-independent Experiments

Finally, to determine the influence of topic specificity on stream adaptation, we perform topic-independent experiments by combining the data across topics. The single-source, multi-source and voting model performances are shown in Table 5. Consistent with our earlier conclusions, we see that the best performance (in bold) is usually achieved either by the native model, or model using all three sources (3-source mixed or voting). However, the benefits of adapted models are not as pronounced as with topic-specific classifiers. For instance, the accuracy of negative classifiers targeting reviews is not matched by any adapted models. This is not true for topic-specific ones, with 3-source mixed models matching native classifier for all individual topics. The same is true for the positive classifiers targeting Twitter.

It is curious, however, that in some of the topic-independent experiments the foreign models significantly outperform the native ones, such as in the case of negative classifiers targeting Twitter (in Accuracy) and targeting reviews (in F-score). Thus, for topically-mixed collections, it is the case that information from a variety of topics from several sources may improve native classifiers. This was not the case for topic-specific experiments earlier, with only one of the multi-source experiments outperforming their native counterparts. We conclude, then, that it is not only beneficial to combine sources of data, but also the topical domains.

Discussion

The experiments described in this work demonstrate the effectiveness of using sentiment classifiers trained on one data source and applied to another. Not only are models trained on single sources often an adequate substitute for the native classifier, but in combination they are even more helpful, often performing as well as the native classifiers. It is interesting that the dataset which was the most challenging to collect and label – the blog stream – was most amenable to classifiers built from other sources. And the dataset which was the least challenging to gather and which did not even need human labeling – the review stream – proved to be the best source of training material for the classification models. It may be the case that the quality of data reviews provide, as well as unambiguous purpose of reviews (that is, to express opinions), overshadow any special language and style features of the other streams. These results are also in agreement with (Bermingham and Smeaton 2010) who find that blogs are the most difficult to classify, followed by microblogs (such as Twitter), and the best classification performance is achieved by models trained on reviews (though note that their work was not on cross-stream classifier experiments).

On the other hand, other streams are not to be discarded in favor of reviews. Twitter is our second best source of

Table 5: Topic-independent source adaptation results. Native classifiers are underlined, best in **bold**, same as native marked with *, better than native marked with †.

		Accuracy						Target F-score					
		Single-source			Mixed			Single-source			Mixed		
Classifier	**Target**	**Blogs**	**Reviews**	**Twitter**	**2 source**	**3 source**	**Voting**	**Blogs**	**Reviews**	**Twitter**	**2 source**	**3 source**	**Voting**
NEG	Blogs	**0.817**	0.732	0.773	0.788	0.801*	0.804*	0.232	0.282*	**0.325**†	0.202*	0.273*	0.256*
	Reviews	0.662	**0.768**	0.642	0.636	0.735	0.715	0.230	0.446	0.434*	0.304	**0.522**†	0.396
	Twitter	0.791	0.762	0.862	0.816	**0.883**†	0.852*	0.141	0.311	**0.450**	0.251	0.354	0.352
POS	Blogs	0.628	0.683†	0.623*	0.660*	0.659*	**0.688**†	0.743	**0.835**†	0.721*	0.752*	0.747*	0.811†
	Reviews	0.790	**0.881**	0.873*	0.821	0.881*	0.878*	0.880*	0.938	0.934*	0.901	**0.939***	0.937*
	Twitter	0.620	0.448	**0.692**	0.631	0.654	0.655	0.565*	**0.669***	0.668	0.551	0.484	0.652*

training data. Unlike reviews, though, it is a much more topically diverse source of data. If one plans on classifying documents about products and services, reviews would be very helpful in building a classifier. But if one is interested in matters outside popular review websites – global issues in policy and economics, or personal ones like self-esteem or social anxiety – reviews may be of little help. It would be interesting to create a multi-dimensional dataset similar to the one in this study, but centered around topical categories not found on popular review websites. We will explore this in future.

Conclusion

In this study we create a multi-dimensional dataset in which three social media sources are queried using a common set of topics and we examine the differences and similarities of the sentiment expressed in these data streams.

Our stream adaptation experiments show the usefulness of each stream as a source of training data. Classifiers built using reviews prove to be the most generalizable to other streams, followed by Twitter, with Twitter-based model performing as well as the native classifier 8 out of 10 for blogs and 5 out of 10 for reviews. We also show that combining training data from several streams further boosts performance, and combining data from different topics may even produce classifiers outperforming their native counterparts.

Our study of the relative usefulness of social media streams as sources of training data allows for more informed design of sentiment analysis tools wherein resources are spent on collecting and labeling data best suited for the task.

References

Asur, S., and Huberman, B. A. 2010. Predicting the future with social media. *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*.

Bermingham, A., and Smeaton, A. 2010. Classifying sentiment in microblogs: Is brevity an advantage? *The ACM Conference on Information and Knowledge Management (CIKM)*.

Blitzer, J.; Dredze, M.; and Pereira, F. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. *ACL* 440–447.

Chesley, P.; Vincent, B.; Xu, L.; and Srihari, R. K. 2006. Using verbs and adjectives to automatically classify blog sentiment. *AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs*.

Chew, C., and Eysenbach, G. 2011. Pandemics in the age of twitter: Content analysis of tweets during the 2009 h1n1 outbreak. *PLoS ONE* 5(11).

Daume, H. 2007. Frustatingly easy domain adaptation. *Association for Computational Linguistics (ACL)*.

Davidov, D.; O., T.; and Rappoport, A. 2010. Enhanced sentiment learning using twitter hashtags and smileys. *COLING*.

Denecke, K. 2008. Using sentiwordnet for multilingual sentiment analysis. *Data Engineering Workshop, ICDEW* 507 – 512.

Kaji, N., and Kitsuregawa, M. 2007. Building lexicon for sentiment analysis from massive collection of html documents building lexicon for sentiment analysis from massive collection of html documents. *EMNLP*.

Kouloumpis, E.; T., W.; and Moore, J. 2011. Twitter sentiment analysis: The good the bad and the omg! *ICWSM*.

Mansour, Y.; Mohri, M.; and Rostamizadeh, A. 2008. Domain adaptation with multiple sources. *Advances in Neural Information Processing Systems (NIPS)*.

Melville, P.; Gryc, W.; and Lawrence, R. D. 2009. Sentiment analysis of blogs by combining lexical knowledge with text classification. *KDD*.

Pan, S.; Ni, X.; Sun, J.-T.; Yang, Y.; and Chen, Z. 2010. Cross-domain sentiment classification via spectral feature alignment. *World Wide Web Conference (WWW)*.

Pang, B., and Lee, L. 2002. Thumbs up?: sentiment classification using machine learning techniques. *EMNLP* 10:79–86.

Paul, M., and Girju, R. 2009. Cross-cultural analysis of blogs and forums with mixed-collection topic models. *EMNLP*.

Peddinti, V. M. K., and Chintalapoodi, P. 2011. Domain adaptation in sentiment analysis of twitter. *Analyzing Microtext Workshop, AAAI*.

Sheng, V.; Provost, F.; and Ipeirotis, P. 2008. Get another label? improving data quality and data mining using multiple, noisy labelers. *KDD*.

Snow, R.; O'Connor, B.; Jurafsky, D.; and Ng, A. 2008. Cheap and fast - but is it good? evaluating non-expert annotations for natural language tasks. *EMNLP*.

Tan, S.; Cheng, Z.; Wang, Y.; and Xu, H. 2009. Adapting naive bayes to domain adaptation for sentiment analysis. *Advances in Information Retrieval* 5478:337–349.

Wan, X. 2009. Co-training for cross-lingual sentiment classification. *Association for Computational Linguistics (ACL)*.

Ye, Q.; Zhang, Z.; and Law, R. 2009. Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. *Expert Systems with Applications* 36:6527–6535.