A Survey and Comparative Study of Tweet Sentiment Analysis via Semi-Supervised Learning

Article *in* ACM Computing Surveys · June 2016

DOI: 10.1145/2932708

3 authors:


Nadia Felix
Universidade Federal de Goiás


Luiz F. S. Coletta
São Paulo State University


Eduardo R Hruschka
University of São Paulo

A Survey and Comparative Study of Tweet Sentiment Analysis via Semi-Supervised Learning

Nadia Felix F. da Silva (nadia@icmc.usp.br)
Luiz F. S. Coletta (luizfsc@icmc.usp.br)
Eduardo R. Hruschka (erh@icmc.usp.br)

August 3, 2016

Abstract

Twitter is a microblogging platform in which users can post status messages, called tweets, to their friends. It has provided an enormous dataset of the so-called sentiments, whose classification can take place through supervised learning. To build supervised learning models, classification algorithms require a set of representative labeled data. However, labeled data are usually difficult and expensive to obtain, which motivates the interest in semi-supervised learning. This type of learning uses unlabeled data to complement the information provided by the labeled data in the training process; therefore, it is particularly useful in applications including tweet sentiment analysis, where a huge quantity of unlabeled data is accessible. Semi-supervised learning for tweet sentiment analysis, although appealing, is relatively new. We provide a comprehensive survey of semi-supervised approaches applied to tweet classification. Such approaches consist of graph-based, wrapper-based, and topic-based methods. A comparative study of algorithms based on self-training, co-training, topic modeling, and distant supervision highlights their biases and sheds light on aspects that the practitioner should consider in real-world applications.

1 Introduction

An increasing amount of content derived from social networking platforms, such as blogs, forums, and microblogs has been observed [104]. Twitter is a famous microblogging service that enable users to post status messages called “tweets” with no more than 140 characters. Tweets represent one of the biggest and most changging datasets of user generated content, with approximately 288 million active users posting 500 million tweets per day[1]. These short texts can express opinions on different topics, which can help to direct marketing campaigns because consumers share their opinions concerning brands and products [37]. Outside the realm of business applications, tweets can make it possible to identify bullying outbreaks [107], events that generate insecurity [12], and acceptance or rejection of politicians [23], all using an electronic word-of-mouth method. Given the huge amount of data that is typically available in the outlined scenarios, actionable insight can be derived from human-machine systems, in which both human expertise and data driven approaches are intelligently combined. To intelligently combine the data, particularly considering our application scenario, four relevant issues must be addressed. Specifically, these issues are:

- Although a tweet can have up to 140 characters, people tend to use much less than this limit. Indeed, the average length of a tweet is 28 characters[2]. This characteristic makes the analyses of tweets based on the so-called bag-of-words harder to perform because the data matrix is very sparse.

- The frequency of misspellings and slang in tweet messages is much higher than that in other domains because users typically post messages from many different electronic devices, such as cell phones and tablets [80]. Furthermore, in this type of environment, users develop their own culture with a specific vocabulary. From the perspective of length, although the content (e.g., in characters) is limited, a message may convey rich meanings.

- Unlike blogs, news, and other sites that are tailored to specific topics, Twitter users post messages on a variety of topics.

- Most tweet sentiment analysis techniques fall into two approach categories: lexicon-based and corpus-based. As with all supervised tasks, these categories require labeled sentiment data to build a machine learning model [83] and/or need labeled sentiment data for evaluation. The more labeled sentiment data that are available, the more robust the machine learning model and the more accurate the evaluation scores.

Our work focuses on the development of tools for tweet sentiment analysis, where labeled data are typically scarce. In this scenario, particular attention

[1]https://about.twitter.com/company

[2]http://thenextweb.com/twitter/2012/01/07/interesting-fact-most-tweets-posted-are-approximately-30-characters-long/

must be given to the role of the (human) experts who help build, monitor, and maintain the system. However, although manual annotation is necessary, it is tedious, expensive, and error-prone [113, 11]. In [27] the authors suggested obtaining labels from emoticons and hashtags, but noted that these are not part of every tweet. Therefore, this and other related approaches [60, 105, 70, 88, 21] have limited use in practice.

Semi-Supervised Learning (SSL) techniques take advantage of using unlabeled data in their training processes and are able to improve classification in applications where labeled data are scarce [2, 28, 115]. In this context, SSL-based approaches show promise in dealing with tweet sentiment analysis because an overwhelming number of unannotated tweets is accessible, in contrast to the limited number of annotated ones [106, 4, 3, 50]. The acquisition of labeled tweets often requires a costly process that involves skilled experts, whereas the acquisition of unlabeled ones is relatively inexpensive. From this perspective, systems based on SSL are of great practical value.

This paper provides a survey of SSL approaches for tweet sentiment analysis. Furthermore, the comparative study conducted offers instructive guidelines for users (experts) interested in practical applications. This type of study is not available in the literature. In contrast to the work of [95], which presumes plenty of labeled data, our work focuses on scenarios where labeled data is scarce. Our work also differs from others that address general approaches for sentiment analysis [56, 26, 98, 48, 62, 102]. In particular, our comparative study considers self-training, co-training, topic modeling, and distant supervision. As a complementary contribution, and to better position our work with respect to the existing literature, we also provide a compact overview of unsupervised and supervised approaches for tweet sentiment analysis.

Our paper is organized as follows. In Section 2, we give a brief overview of the literature on supervised and unsupervised approaches for tweet sentiment analysis. This overview describes the detailed and systematic survey of SSL approaches in Section 3. In Section 4, we report an experimental comparative analysis performed on representative approaches that are surveyed in Section 3. Finally, in Section 5, we summarize our study and conclusions as well as we address some important issues for future research.

2 A Brief Overview of Supervised and Unsupervised Sentiment Analysis

Most of the studies about tweet sentiment analysis utilize supervised learning algorithms to produce sentiment classification models (see Figure 1). Such algorithms require a training set formed by labeled data, where the labels are the classes (e.g., positive, neutral, and negative) of each tweet. Some studies propose the use of emoticons and hashtags for building the training set, including [27] and [21], who identified tweet polarity by using emoticons as class labels — this type of strategy is known as distant supervision based classification. Deep learning approaches have also used emoticons as class labels to refine the embeddings on a large distant supervised corpus [84, 91, 94, 92, 93]. Other algorithms use the characteristics of the social network as networked data, as in [36].

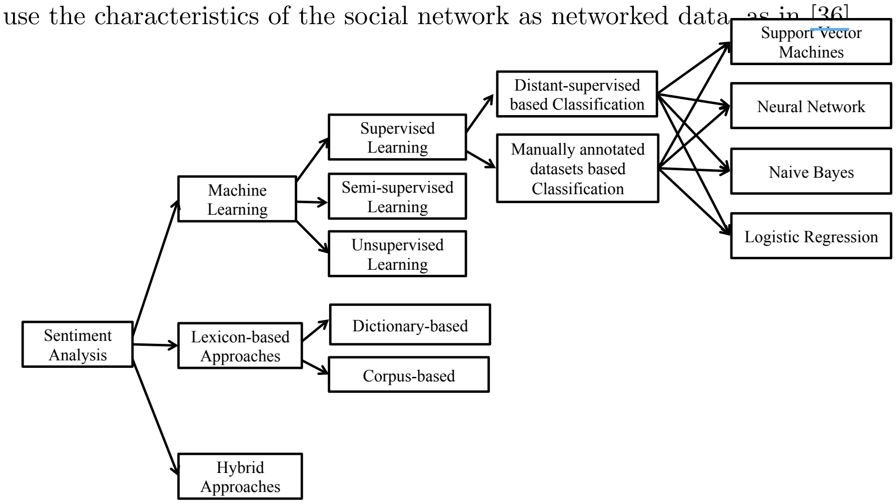


Figure 1: Overview of tweet sentiment analysis approaches.

The lexicon-based approaches depend on the availability of a sentiment lexicon, which is a collection of known and previously created sentiment words. These approaches can be categorized into two different groups: (i) Dictionary-based, which use dictionaries as lexical resources [89, 71, 35, 42], and (ii) Corpus-based, which use statistical or semantic methods to find sentiment polarity [100, 99].

Approaches that integrate opinion mining lexicon-based techniques and machine learning-based techniques have also been investigated (known as hybrid approaches). For example, [1, 74, 110, 61] used lexicons, part-of-speech, and writing style as linguistic resources. In a similar context, [79] introduced an approach to add semantics to the training set as an additional feature. More recently, classifier ensembles have been successfully used [18, 86, 47, 14, 77, 33].

The seminal work on sentiment classification that does not depend on labeled data was proposed by [100], in which a document is predicted as either positive or negative by taking into account the semantic orientation of its phrases that contain adjectives or adverbs. His approach was assessed on automobile reviews and movie reviews, which are data sources that are very different from short texts such as those found in tweets. Along the same line, [75] put forward different forms to quantify the similarity between words and polarity words (based on lexical association, semantic spaces, and distributional similarity). Because labeled data are not used by unsupervised learning approaches, they are expected to be less accurate than those based on supervised learning. From this aspect, prominent human-machine systems for sentiment analysis should address the scarcity of labeled tweets (taking advantage of unlabeled ones, such as is done by unsupervised models) and provide better classification (as those usually reached by supervised models).

3 Semi-supervised Learning for Sentiment Analysis

Semi-Supervised Learning takes advantage of both unlabeled and labeled data during the training phase [2, 28, 115]. Therefore, as shown in Figure 2, SSL fits in between supervised and unsupervised learning. For supervised approaches, all training instances must be labeled and more interactivity with the users (experts) is required. This dependency decreases in SSL approaches, in which a balance between supervised and unsupervised learning is found [11].

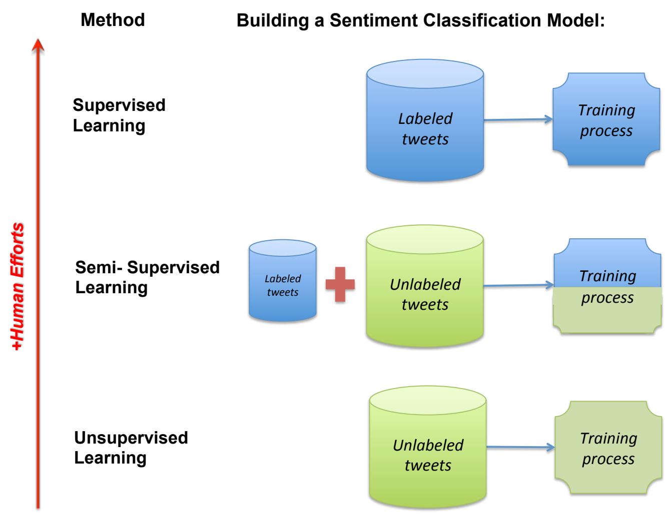


Figure 2: Typical methods of learning according to required human efforts. Semi-Supervised Learning establishes a synergy between supervised and unsupervised learning by compensating for the lack of labeled instances with unlabeled ones and is thus particularly useful for building sentiment classification models.

Given a labeled data set, $D_l = \{(x_i,y_i)|(x_i, y_i) \in X \times Y, i = 1, ..., l\}$, and an unlabeled data set, $D_u = \{x_j|x_j \in X, j = l+1,..., l+u\}$, in which X denotes the input space of data instances and Y is the label space, according to [116] "a semi-supervised algorithm aims to train a classifier f from $D_l \cup D_u$, i.e., from both the labeled and unlabeled data, such that it is better than a supervised learner induced on the labeled data alone" (p.9). Thus, SSL is particularly appropriate in cases where obtaining an unlabeled sample is cheap and easy, while labeling the sample is expensive or difficult [11] — as a consequence, typically unlabeled data is much more accessible and available than labeled data, i.e., $u \gg l$. This is the case of several sentiment analysis applications, especially when the data source comes from social networks (e.g. Twitter).

We identify three categories of semi-supervised approaches for tweet sentiment analysis: (i) graph-based methods, (ii) wrapper-based methods (e.g., self-training and co-training), and (iii) topic-based methods. To address these categories and understand the context and the development of research on the subject, we also provide an overview of approaches that manipulate other types of data sources, including web pages, online news, Internet discussion groups, online reviews, and web blogs.

3.1 Graph-based Methods

Graph-based methods propagate labels to unlabeled data. The label propagation process requires the computation of similarities among the data instances. Similarities are captured through a graph $\mathcal{G} = \langle \mathcal{V}, E \rangle$, where each vertex v_i from the vertex set $\mathcal{V}$ represents an instance $x_i \in X$ and each edge (v_i, v_j) from the edge set E is associated with a non-negative weight w_{ij}. Such a weight indicates the similarity between v_i and v_j. In addition, $\mathcal{V} = \mathcal{V}_l \bigcup \mathcal{V}_u$, where each vertex in $\mathcal{V}_l$ has an initial label $y \in Y$ and all vertices in $\mathcal{V}_u$ are unlabeled.

In the area of sentiment analysis, the literature on graph-based algorithms focuses on sentiments related to either sentences or full documents. The investigated applications range from document polarity classification [81, 87, 76], document rating prediction [29, 114], and identification of political affiliation [53] to algorithms that learn sentiment polarity lexicons from a few seed words [73].

The use of a suitable similarity measure, usually dependent upon the specific task of interest, is the key to the successful application of graph-based algorithms because it determines the distance between two data instances and, as a consequence, how similar the probability distributions of their labels should be. In other words, such algorithms work only if a proper similarity measure exists such that the assumption holds. For document-level sentiment analysis, finding the similarity measure is non-trivial. Typically, cosine similarity based on bag-of-words representation is employed. However, this favors topic similarity rather than sentiment similarity. Indeed, a high similarity value often suggests that two documents share numerous content words rather than similar sentiments. As shown in [29], using the cosine similarity with bag-of-words representation, algorithms performed worse than the support vector regression [39] for rating prediction of movie reviews. By changing that measure to the positive-sentence percentage-based similarity, which is computed as the percentage of positive sentences in a document, graph-based algorithms outperform their supervised counterparts when the quantity of labeled documents is limited [67].

Instead of defining a proper similarity measure to construct a similarity graph, [81] and [87] proposed different methods for constructing similarity graphs. [81] encoded prior knowledge into a graph of word features, in which the vertices represent words and the edges represent similarities between them.

On the other hand, the use of graph-based methods has been motivated by the available social information, which can help to capture sentiments of particular users [90, 69]. Users that can somehow be categorized as “follower” or “followee” are more likely to hold similar sentiments. Accordingly, relationship information can help what can be extracted about users perspectives that are originated from textual features only. It is worth noting that similarity measures still have a key role in these approaches. However, they are now based on users characteristics. In particular, the graph should capture the fact that some users share similar opinions. These approaches are supported by many social studies [44, 55, 96].

[90] proposed a formulation of a “Twitter Graph”, where it is considered a “query” topic that includes users who have tweeted about this, while omitting users who have never expressed themselves about the query topic. The goal is to distinguish between users that show positive feelings about the topic and those who have negative feelings about the topic. A connection edge between two users is set up if one follows or mentions the other.

[40] investigated a graph based method called “label propagation”. The general idea behind their method is to build a weighted graph where the users, tweets, and other features are the set of vertices. The edges connecting the vertices are derived from retweets and their weights are related to the relative frequency ratio of the unigram or bigram in the training data. Given such a graph structure, a label distribution is initially seeded to a subset of vertices and then spread across the graph.

[9] proposed a transfer learning approach for tweet sentiment analysis. It is based on textual resources and the prediction of social media user bias. Transfer learning is applicable when classification is hampered (e.g. because of outdated data and lack of labeled instances) and improvements can be reached through supplementary knowledge that is derived from similar concepts [46, 66]. According to [9], it is possible to use “social media endorsements from retweets to quantify user bias towards a topic. Endorsements may be represented as a directed graph, where an edge represents that a user endorsed or retweeted a tweet from a user”.

3.2 Wrapper-based Methods

A wrapper-based method uses a supervised learning algorithm in an iterative fashion. In each iteration, a certain amount of unlabeled instances is labeled by the decision function that is learned and incorporated into the training data. From its own predictions and the labeled data already available, the classification model is retrained for the next iteration. The well-known representatives of this category are self-training [82] and co-training [8]:

3.2.1 Self-training

Essentially, [115] state that “the learning process uses its own predictions to teach itself” (p. 15). Self-training starts with a supervised learner that is trained on available labeled data and then iterates several times. In each iteration, it selects a subset of predicted instances to augment the training data. Typically, this subset contains instances for which the predictions have shown higher confidence levels. Then, the new training data are used to update or retrain the supervised learner for the next iteration. This iterative learning process implies that the method only works if the highest confidence predictions are effectively correct [116].

Self-training has been applied in several contexts. For example, the algorithm AROW [16] makes use of self-training for large scale reviews of polarity prediction. [32] show that AROW can reduce test errors by more than half compared to the supervised classifier trained on the initial labeled data.

Another sentiment classification approach based on self-training was proposed by [109]. In this work, self-training is used to add sentiment lexical items into the vocabulary for Chinese text. [51] also used a self-training approach for sentiment analysis in a Chinese microblog.

Similarly to [109], [72] utilized an iterative process based on lexicon to increase a sentiment dictionary. The approach considers a massive Chinese sentiment dictionary instead of employing a one-word seed dictionary as in [109]. Documents predicted in the initial phase are used as the training data to build Support Vector Machines (SVMs), which are subsequently employed in the refinement of the primary results.

Finally, approaches that employ self-training for increasing the size of the feature space can be found in [4, 3, 111], in which the training process leads to the inclusion of additional polarity lexicons. The main motivation of these approaches is to adapt a static polarity lexicon with the help of an unlabeled tweet set.

3.2.2 Co-training

It adopts an iterative learning process similar to self-training, but instead of using a single supervised learner, it uses two learners that teach each other [8]. The two learners operate on different feature sets, which are referred to as (independent) views of the data. The most confident classifications from each learner are then used to iteratively build more labeled data which, in turn, are used for training. The process finishes when all unlabeled data have been used or a specific number of iterations has been reached.

As in self-training, co-training has successful applications in sentiment analysis. For the cross-lingual document polarity classification [103], each view used by co-training is a set of language-based features. For problems where the class distribution is imbalanced, [45] proposed an under-sampling method to generate balanced datasets of different views. [108] revisited co-training in depth, discussing several strategies for sentiment analysis in three domains: news articles, online reviews, and blog posts. [50] also designed a two-view (textual and non-textual views) approach for tweets classification based on the co-training framework. In their approach, two classifiers are trained on a common set of labeled tweets. According to [49], they proposed a semi-supervised adaptive

SVM model that augments the labeled set and expands topic-adaptive features based on the unlabeled data available.

3.3 Topic-based Methods

The methods reviewed above only consider features that capture local information in the data (e.g., lexicons, unigrams[3], bigrams[4], part-of-speech, etc.). Specifically, they do not consider global, higher-level information, such as topic information that may somehow influence sentiments. In particular, the same word may have different sentiment polarities in different domains. For instance, though the adjective "*complex*" in the sentence "*The book is complex and exciting!*" may have a positive orientation in a book review, it could also have a negative orientation in the sentence "*It is hard to use such a complex cellphone*" in an electronic review. Therefore, it is more suitable to analyze topics and sentiments simultaneously.

Topic information has been applied in different domains of sentiment analysis. In the seminal studies [57, 38, 34], all of the training data are required to infer the classes of unlabeled instances. More recently, [85] used a continuous Dirichlet Process Mixture model to learn daily topic sets. Then, for each topic, the sentiment is derived according to an opinion word distribution aiming to build a sentiment time series. The sentiment was estimated based on a lexicon (a list of positive and negative opinion words, e.g., "*good*" and "*bad*").

In [106], a topic-based SSL is used to analyze sentiments from tweets. The authors proposed building a topic model on labeled tweets[5] so that specific sentiment models can be induced on each cluster that was found. Figure 3 summarizes such an approach. Firstly, a classifier is inferred from labeled tweets, then it is used to estimate the class probabilities for each unlabeled tweet (this primary step occurs only once). After this (in Step 4), a subset of tweets with class probability higher than a confidence threshold is selected. They are then included in the labeled tweets set. In step 5, the labeled tweets are used to build a topic model, from which topic distributions are stored for each tweet. Then, clusters based on the topic distributions are inferred and a particular sentiment model is trained for each cluster. The resulting sentiment mixture model is used to classify the unlabeled tweets (Steps 7 and 3). An iterative process takes place until a certain number of iterations has been reached or no more tweets have been promoted to the set of labeled tweets.

4 Comparative Study

A comprehensive comparison on SSL methods for tweet sentiment analysis is not an easy task. The main difficulties come from the fact that there is no

[3] Words or tokens.

[4] Sequence of two adjacent words in a text of tokens.

[5] The topic information is generated through topic modeling based on the implementation of Latent Dirichlet Allocation (LDA) [7].

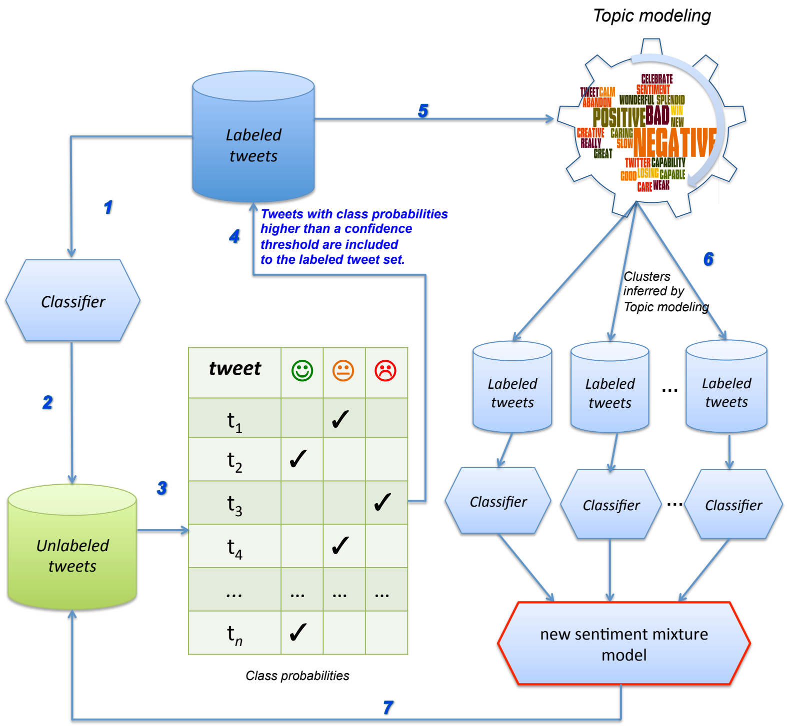


Figure 3: Topic-based approach for tweet sentiment analysis proposed in [106].

consensus about which features are the best or which proportion of unlabeled data should be used. Most of the datasets have limited use because they are not publicly available or because they involve proprietary data.

Our goal was to conduct controlled experiments with fair and instructive comparisons among the different methods[6]. To do so, we used a set of standardized features from public data, using no additional data to formulate the semi-supervised phase. Table I shows an overview of the studies on SSL for tweet sentiment analysis that was surveyed in Section 3. These studies employ evaluations performed on proprietary (unlabeled) data; thus, reproducibility is obviously an issue. In addition, we did not include graph-based approaches because they require information about the user network.

4.1 Datasets

Table 2 summarizes the datasets used in our comparative study. We utilize datasets employed by the organizers of the International Workshop on Seman-

[6]Software is available upon request from the authors.

Table 1: Overview of literature on SSL for tweet sentiment analysis. Evaluations were carried out on proprietary unlabeled data though the training data were typically public. F-Scores are shown for illustrative purposes.

Approach	**Work**	**Dataset**	**#labeled**	**#unlabeled**	**F-score**
Self-training	[4]	SemEval 2013 [63]	9,829 – Public	485,112 – Proprietary	0.641
	[3]	SemEval 2013 [63]	8,750 – Public	910,000 – Proprietary	0.543
	[111]	SemEval 2013 [63]	8,471 – Public	N/A – Proprietary	0.637
Co-training	[50]	TREC 2011 Microblogging and proprietary dataset	16,000,000 – Proprietary	N/A – Proprietary	N/A
	[49]	"Taco Bell", Sanders-Twitter Sentiment, and 2008 Presidential debate corpus (all together)	10,537 – Public	N/A – Proprietary	The performance was assessed on different sample ratios
Topic Modeling	[106]	SemEval 2013 [63]	9,684 – Public	2,000,000 – Proprietary	0.703

tic Evaluation (SemEval)[7] which is a leading scientific event in this field. As suggested by the organizers of SemEval 2013 (task 2) and SemEval 2014 (task 9) competitions, the dataset known as SemEval 2013 was used to induce classification models. Actually, this dataset is currently the most used for tweet sentiment analysis, in addition to being representative and publicly available with a considerable size [63, 78]. The induced models were then assessed on five test sets, namely LiveJournal, SMS2013, Twitter2013, Twitter2014, and Twitter Sarcasm 2014. Essentially, the datasets LiveJournal and SMS2013 were included to determine how systems that are trained on Twitter perform on other sources (particularly, from web blogs and cell phone messages). They were labeled by the Amazon Mechanical Turk[8] annotators. Twitter2013 was obtained in a process formed by three phases: firstly, named entities were extracted from millions of tweets that were collected over a one-year period spanning from January 2012 to January 2013 using the public streaming Twitter API. Then, popular topics, such as those named entities that were frequently mentioned in association with a specific date, were identified. Finally, given this set of automatically identified topics, tweets were gathered from the same time period related to the named entities. Twitter2013 has different topics from training and spanned later periods. Twitter2014 and Twitter Sarcasm 2014 were obtained more recently. The latter was collected by the #sarcasm hashtag with the goal of determining how sarcasm affects the tweet polarity.

[7]http://en.wikipedia.org/wiki/SemEval
[8]https://www.mturk.com

Table 2: Class distributions of training and test sets that were used. Sentiment classification models were induced on a set of labeled tweets (SemEval 2013 [63]). The results were obtained on five test sets, namely LiveJournal, SMS2013, Twitter2013, Twitter2014, and Twitter Sarcasm 2014.

Training set				
Name	*Positive*	*Negative*	*Neutral*	*Total*
SemEval 2013 [63]	4,215 (37%)	1,807 (15%)	5,325 (48%)	11,338
Test sets				
LiveJournal [78]	427 (37%)	304 (27%)	411 (36%)	1,142
SMS2013 [63]	492 (23%)	394 (19%)	1,207 (58%)	2,093
Twitter2013 [63]	1,572 (41%)	601 (16%)	1,640 (43%)	3,813
Twitter2014 [78]	982 (53%)	202 (11%)	669 (36%)	1,853
Twitter Sarcasm 2014 [78]	33 (38%)	40 (47%)	13 (15%)	86

4.2 Feature Engineering

Different studies have used different features to represent tweet messages. In fact, it is expected that a set of the chosen features properly fits the classification model adopted. For example, approaches in Table 1 have employed Ngrams and emoticons [50, 49] or only Ngrams [3]. Others adopted a more complex feature space also containing part-of-speech tags, lexicons and hashtags [4, 106, 111]. The feature set used in our experiments was inspired from [61], whose authors ranked first in SemEval 2013 [63]. Such feature set also achieved the highest scores on LiveJournal, Twitter Sarcasm 2014, and SMS2013 in SemEval 2014 [78]. It is composed of:

(i) *Ngrams:* unigrams, bigrams, and trigrams.

(ii) *Negation:* the number of negated contexts. A negated context according to [68] is a segment of a tweet that starts with a negative word (e.g., *"no"*, *"shouldn't"*) and ends with a punctuation as a comma, period, colon, semicolon, exclamation mark, or question mark. A negated context affects the ngram and lexicon features, so that it added the suffix *"NEG"* to each word following the negation word (e.g., *"good"* became *"good_NEG"*). A list of negation words was adopted from Christopher Potts' sentiment tutorial[9].

(iii) *Part of Speech:* a part-of-speech tagging was carried out by using Ark-twitter NLP [65] and the number of occurrences of each part-of-speech tag was computed.

(iv) *Writing Style:* we considered the presence of three or more repeated characters in the words, the sequence of three or more punctuation marks, and

[9] http://sentiment.christopherpotts.net/lingstruc.html

the number of words with all letters in uppercase.

(v) *Lexicons:* the number of positive and negative words computed by the lexicon-based method [61]

(vi) *Microblogging features:* the total number of sentiment hashtags in the text provided by sentiment lexicons and emoticons [35, 97, 61].

4.3 Experimental Setup

To perform a fair comparison among the existing semi-supervised tweet sentiment analysis methods, we used only public datasets, as shown in Table 2. The literature indicates that Naive Bayes, SVM with linear kernel, and Logistic Regression are the most used algorithms in tweet sentiment analysis [63]. We have chosen SVM to perform our experiments because it provides a good out-of-sample generalization, usually providing better classification accuracy compared to Naive Bayes and Logistic Regression in tweet classification applications. As in [61], we used a linear kernel with parameter $C = 0.005$.

Two widely known SSL approaches were used in our comparative study, namely self-training [82] and co-training [8]. The advantages and disadvantages of the topic-based approach introduced in [106] are also analyzed.In addition, we performed a tweet sentiment classification using distant supervision [27], where the sentiment classes in the training set are replaced by positive and negative emoticons and hastags, and in cases that there are no emoticons the tweets are considered as neutral. Note that, in this method, human effort is not required to annotate the training set. Thus, this is indeed a useful baseline for comparison purposes. For self-training and distant supervision based classification, we considered all features mentioned in Section 4.2. Because the model being constructed learns iteratively by aggregating new reliable data, we adopted specific confidence thresholds according to [82]. In particular, because the classifier gives confidence scores when it labels instances from unlabeled data, those instances with confidence scores higher than the predefined threshold are promoted to the labeled set. In our experiments, we evaluated the following values of this parameter: 0.4, 0.5, 0.6, 0.7, 0.8, and 0.9. For co-training, we considered the textual features (i.e., unigrams, bigrams, and trigrams) as one view, and the features from ***(ii)*** to ***(vi)*** in Section 4.2 as the second view (also referred as lexicon view). Figure 4 illustrates the use of these views by the co-training approach. Based on [8], we set the number of samples per class, which are classified with the best confidence levels, as $p = 3$, $n = 2$, and $ne = 4$ for positive, negative, and neutral classes, respectively. These parameters were defined from the distribution of classes in the training set. The size of the smaller pool U' was set to 10% of the training set.

To evaluate how the algorithms perform with different amounts of initial labeled data, we randomly sampled a proportion, s, of labeled tweets from the training set (maintaining the balance of the three classes). The remaining

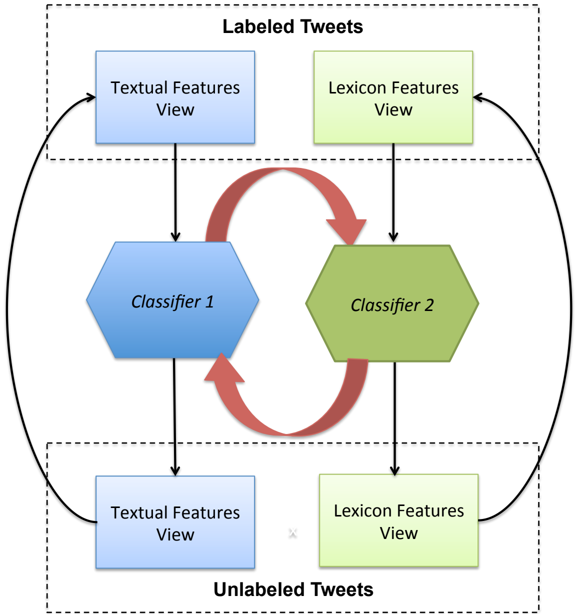


Figure 4: Conceptual schema of textual and lexicon views being used in the co-training method. The classifier models cooperate between themselves.

$(1-s)$ tweets were used in the learning phase — in which a certain number of instances were incorporated and used for adapting the classification model (in our experiments such a phase consisted of 40 iterations of the algorithms). The proportion of initial labeled tweets, s, was varied as a percentage of the number of instances from the training set as 1%, 5%, 10%, 20%, and 40%.

The F-score ($F1$) was adopted for evaluating the accuracy of the algorithms. We computed the F-score for each class (positive, negative, and neutral) and the overall F-score ($\overline{F1}$), which was obtained by $(F1_{positive} + F1_{negative})/2$ [4, 3, 111, 106]. Algorithms were run 20 times and the averages and standard deviations were reported.

The computational implementation uses the Natural Language Toolkit (nltk)[10] for preprocessing tweet messages, the Scikit-Learn (sklearn)[11] for classification (SVM), and gensim[12] for topic modeling with hierarchical LDA.

[10] http://www.nltk.org
[11] http://scikit-learn.org/
[12] http://radimrehurek.com/gensim/

4.4 Results and Discussion

4.4.1 Topic-based approach

The performance of the topic-based approach (Section 3.3) is highly dependent on the choice of the confidence threshold and the inference of the number of topics (clusters). If the number of labeled tweets is small and the chosen confidence threshold is high, the learning process will be hampered, whereas if the threshold is low, the model will learn wrong classes. In [106], the authors chose an arbitrary value for the number of topics. We adopt a more principled approach, based on the Hierarchical Dirichlet Process (HDP) [31], which is widely used in applications where different groups of data may share the same settings of partitions. The adopted approach does not require the number of topics to be provided in advance, i.e., the number is estimated directly from data.

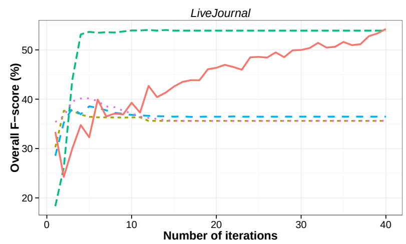
Because the training set was collected over a one-year period, spanning from January 2012 to January 2013, we may have different samplings with a wide variety of tweets. In this scenario, the HDP tends to obtain clusters with tweets from only one class or two classes, especially if the threshold is high. After clustering the training set based on topic distributions, the next phase is to train a separate sentiment model for each cluster. However, if the clusters only have tweets from one or two classes, then the algorithm cannot proceed because the groups of tweets must reflect the probability distribution over the three classes under study.

The algorithm was unable to learn the classes with 1%, 5%, 10%, 20% and 40% of the training set and confidence thresholds ranging from 0.4 to 0.9. Taking into account samplings with 60% of the training set and the threshold set to 0.9, some learning progress was observed, but it is still not compatible with the self-training and co-training algorithms. It is likely that the algorithm worked well in [106] because the authors used 2M tweets as additional unlabeled data, with a threshold of 0.96, and all the training set consisted of labeled tweets. By doing this, it is possible to have representative clusters that, in turn, are good models of classification.

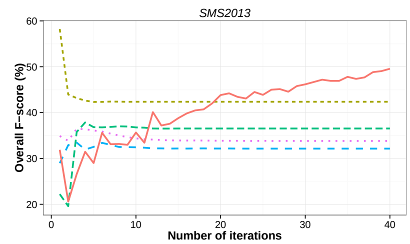
4.4.2 Self-training and co-training

We focus on the overall F-score curves as the number of promoted instances increases over the iterations, as well as on the different amounts of the initial labeled instances. In particular, we explored the F-score relation between positive and negative classes.

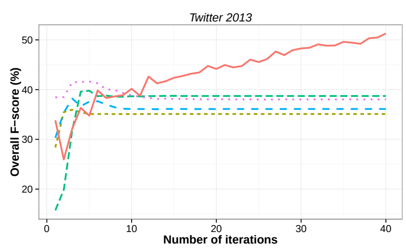
Figure 5 illustrates the overall F-score curves when 1% of the training set — SemEval 2013 [63] — was used as initial labeled data. Algorithms run for 40 iterations. Co-training resulted in significant learning, demonstrated by the increasing F-score curves, particularly for the datasets LiveJournal, SMS2013, Twitter2013, and Twitter2014. For self-training, low confidence thresholds (such as 0.6) are likely to have deteriorated the learning process because the promoted instances are more likely to have been misclassified. However, higher thresholds (such as 0.9) allow selecting more useful and noise-free data, but are typically more difficult to obtain. Self-training with a threshold set at 0.9 achieved good F-scores for Twitter 2014. By setting the threshold at 0.7 (i.e., a more balanced threshold), self-training resulted in a competitive performance on LiveJournal and better results on Twitter Sarcasm 2014. This occurs because LiveJournal has texts from a web blog, where there is less slang and typos, and no character limit for the user.



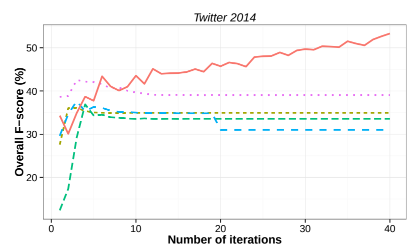
(a)



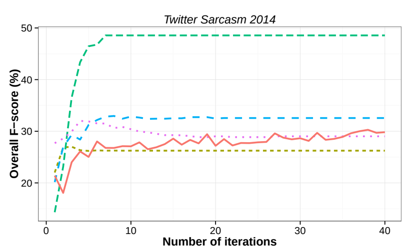
(b)

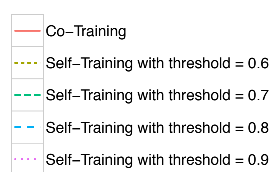


(c)



(d)





(e)

Figure 5: Overall F-score curves for 40 iterations of the self-training and co-training approaches. For self-training, four different confidence thresholds were assessed. The size of initial labeled data corresponds to 1% of the original training sets.

Figure 6 shows the overall F-scores achieved by different proportions of initial labeled tweets (1%, 5%, 10%, 20% and 40%) after 40 iterations of the algorithms. As can be seen, with limited labeled instances (e.g., 1% of the training set) the best choice is to use co-training. However, if more labeled instances are available, self-training can obtain better results, being potentially more useful for these scenarios. Typically, self-training is an algorithm with a sensitive parameter; however, in our experiments we observed that self-training with a confidence threshold equal to 0.9 offered the best results in general (i.e., considering different proportions of initial labeled tweets). It is worth mentioning that in the presence of irony and sarcasm (i.e., considering the Twitter Sarcasm 2014 dataset), self-training was the best choice when the size of initial labeled data was 5%, 10%, 20%, and 40%.

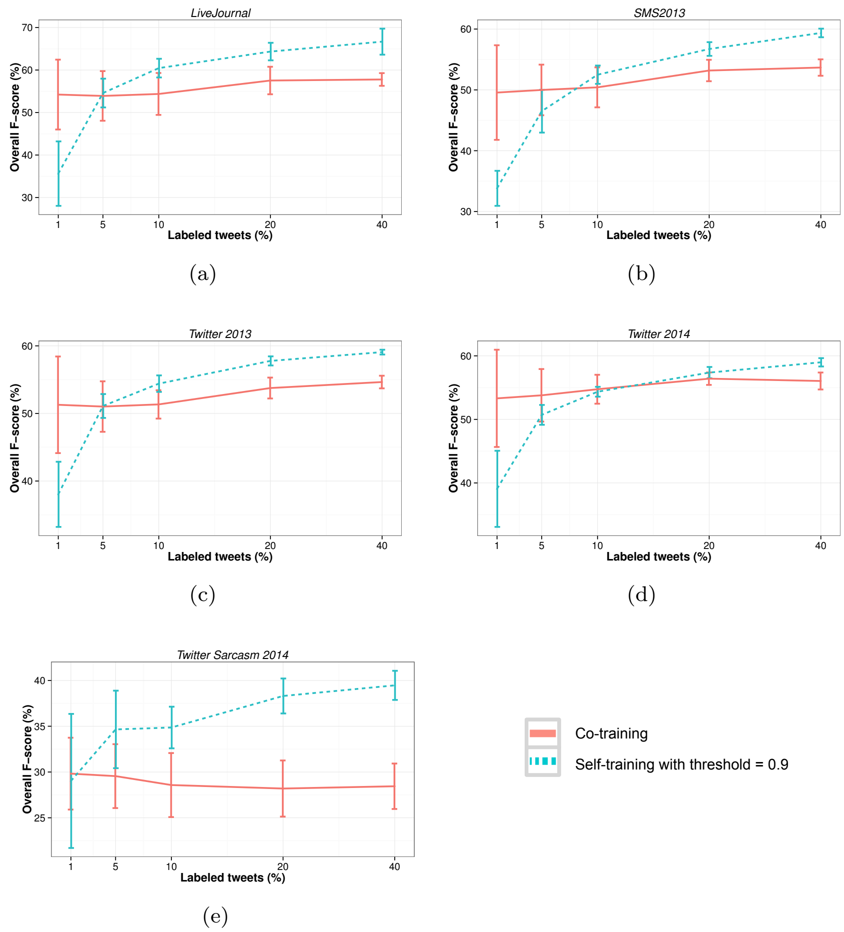


Figure 6: Overall F-score for different sizes of initial labeled data, which correspond to 1%, 5%, 10%, 20%, and 40% of the original training sets. Such results were obtained after 40 iterations of the algorithms.

Figures 7 and 8 show the F-scores for positive and negative classes (individually) after 40 iterations of the algorithms and different amounts of initial labeled instances. These results show that co-training performed better with limited data and without the presence of irony and sarcasm. With at least 10% of the training set as initial data, self-training is the best choice to solve the problem.

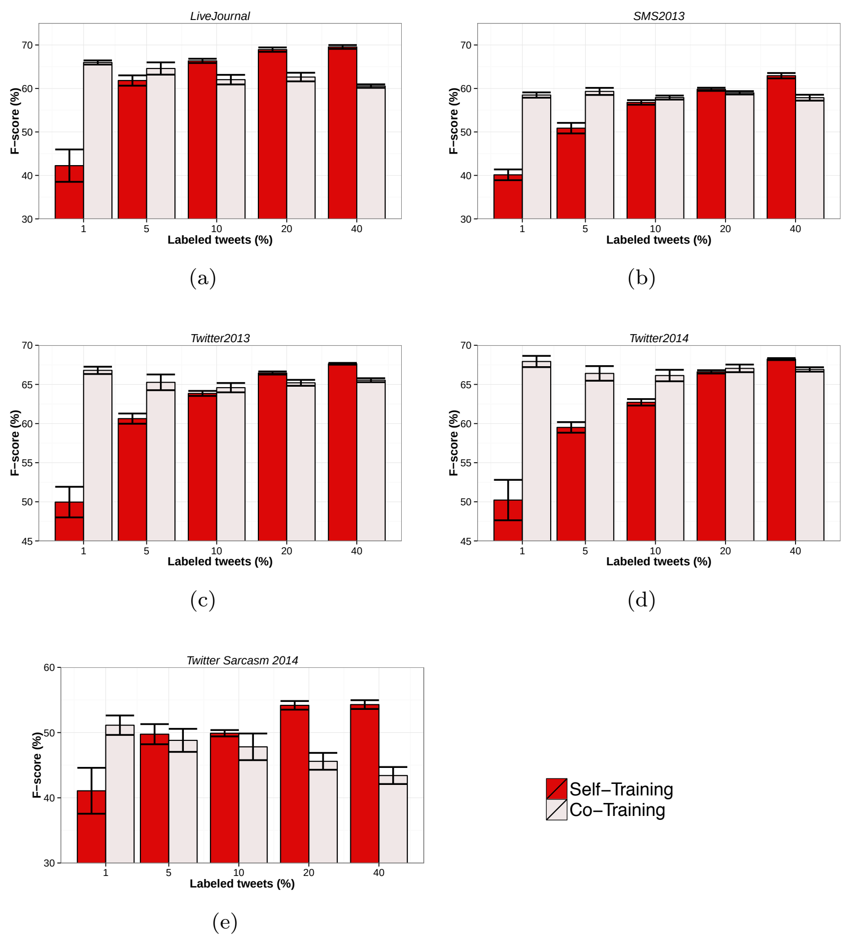
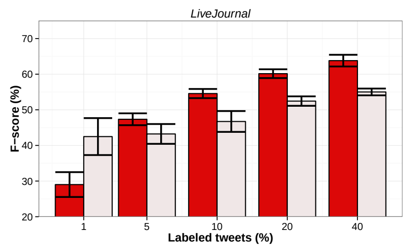
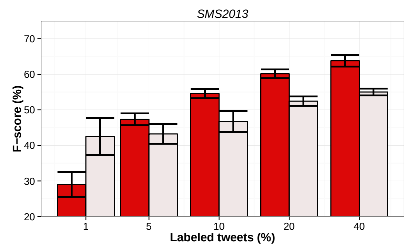


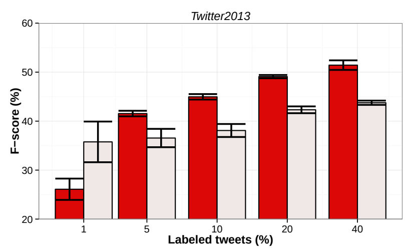
Figure 7: F-scores for the positive class and different percentages of initial labeled data. Such results were obtained after 40 iterations of the algorithms.



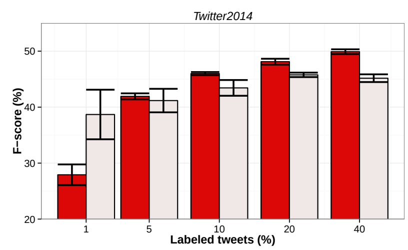
(a)



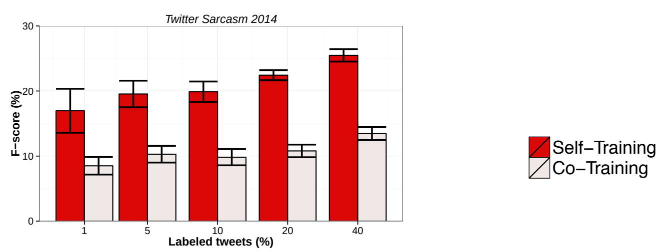
(b)



(c)



(d)



(e)

Figure 8: F-scores for the negative class and different percentages of initial labeled data. Such results were obtained after 40 iterations of the algorithms.

Table 3 shows specific results for the self-training (with confidence threshold of 0.9) and co-training approaches. The F-score measure for each class (positive, negative, and neutral) and the overall F-score, which was obtained by averaging the F-scores from positive and negative classes, are presented. Better results are highlighted in bold face and the best results found in the literature are also reported. From this table, we can extract some interesting findings:

1. Because the LiveJournal set is composed of formal texts with no slang, with 40% of the training data the self-training approach obtained an overall F-score of 66.68%, which is close to the overall F-score achieved by SVM that was implemented with the whole training set (67.34%).

2. Because SMS2013 has a well-defined vocabulary (with known slang), with only 1% of labeled instances, the co-training approach obtained an overall F-score of 52.39%, which is close to the same result that was obtained for the whole training set (55.50%) with SVM. Better results were achieved by using 40% of the training set and self-training, with an overall F-score equal to 59.36%.

3. By using 40% of the training set, an overall F-score of 59.13% was obtained by self-training on Twitter2013. For Twitter2014, in the same scenario, the overall F-score is 58.95%. Such values are close to the results observed when the whole training set is used (66.05% and 63.69%, respectively).

4. [112] achieved the best results on LiveJournal, SMS2013, and Twitter Sarcasm 2014 by training a SVM classifier with all labeled instances available and using cluster features as extra features that were inferred from 56 million English language tweets. The 1,000 clusters found are an alternative representation of tweet content. The authors emphasized that this strategy reduces the sparsity of the token space, because the n-grams-based features are replaced by the representative elements of the data partition.

5. [59] yielded the best results on Twitter2013 and Twitter2014 by using Logistic Regression on all labeled data. Several lexicons and pre-processors were employed to enhance the lexical information. In addition, because the distribution of sentiment on training set is previously known, the authors proposed a weighting scheme that biases the learning process.

6. Although self-training with 40% of the training set provided competitive results (39.65%) in comparison with SVM on all labeled tweets (41.08%), the modest results on Twitter Sarcasm 2014 suggest that more research efforts are necessary. In particular, the study of features that can properly represent irony and sarcasm is required. Such features may leverage the performance of human-machine systems in these scenarios. It is also possible that by using a different evaluation measure (other than F-score), one might get more competitive results on Twitter Sarcasm 2014.

7. Table 4 summarizes the results for self-training and co-training by providing the percentages of higher F-scores and lower standard deviations found

Table 3: F-scores for different percentages of the initial labeled data for self-training, co-training, and supervised approaches from the literature.

LiveJournal

Approach	1%				5%				10%				20%				40%				100%			
	F1-pos	F1-neg	F1-neu	$\overline{F1}$	F1-pos	F1-neg	F1-neu	$\overline{F1}$	F1-pos	F1-neg	F1-neu	$\overline{F1}$	F1-pos	F1-neg	F1-neu	$\overline{F1}$	F1-pos	F1-neg	F1-neu	$\overline{F1}$	F1-pos	F1-neg	F1-neu	$\overline{F1}$
Self-training	42.24 ±11.78	29.02 ±11.03	50.68 ±14.81	35.63 ±7.58	61.82 ±3.74	**47.35 ±5.25**	62.39 ±4.68	**54.59 ±3.38**	**66.34 ±1.54**	**54.57 ±4.06**	**66.80 ±0.75**	**60.45 ±2.19**	**69.39 ±0.87**	**60.15 ±3.90**	**68.62 ±0.92**	**64.77 ±2.03**	**69.54 ±1.39**	**63.81 ±5.18**	**70.26 ±0.73**	**66.68 ±3.06**	–	–	–	–
Co-training	**65.97 ±1.51**	**42.49 ±16.40**	**63.28 ±5.28**	**54.23 ±8.21**	**64.59 ±4.46**	43.23 ±8.79	**64.99 ±1.35**	53.91 ±5.84	62.02 ±3.49	46.73 ±9.25	64.19 ±1.41	54.38 ±4.93	62.62 ±3.13	52.45 ±4.19	65.55 ±0.85	57.54 ±3.23	60.54 ±1.28	55.02 ±3.01	65.71 ±0.78	57.78 ±1.49	–	–	–	–
SVM	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	68.84	61.75	55.37	67.34
Literature	Supervised approach proposed by Zhu et al. [112]																							74.84

SMS2013

Approach	1%				5%				10%				20%				40%				100%			
	F1-pos	F1-neg	F1-neu	$\overline{F1}$	F1-pos	F1-neg	F1-neu	$\overline{F1}$	F1-pos	F1-neg	F1-neu	$\overline{F1}$	F1-pos	F1-neg	F1-neu	$\overline{F1}$	F1-pos	F1-neg	F1-neu	$\overline{F1}$	F1-pos	F1-neg	F1-neu	$\overline{F1}$
Self-training	39.48 ±3.71	28.15 ±4.38	54.72 ±20.28	33.81 ±2.88	50.87 ±3.83	**42.13 ±4.18**	64.81 ±7.55	46.50 ±3.51	56.78 ±1.67	**48.21 ±2.59**	72.18 ±1.75	**52.50 ±1.50**	**59.80 ±1.15**	**53.64 ±1.73**	75.60 ±1.75	**56.72 ±1.12**	**62.93 ±1.95**	**55.79 ±2.73**	76.28 ±4.20	**59.36 ±0.70**	–	–	–	–
Co-training	**59.23 ±3.17**	**43.13 ±17.28**	**75.93 ±2.97**	**52.39 ±12.14**	**59.32 ±2.57**	40.66 ±7.07	**79.03 ±0.75**	**49.99 ±4.16**	**57.90 ±1.47**	42.93 ±6.48	**79.53 ±0.63**	50.42 ±3.28	59.00 ±1.21	47.38 ±3.13	**80.13 ±0.37**	53.19 ±1.76	57.88 ±2.12	49.47 ±1.40	**80.56 ±0.23**	53.67 ±1.35	–	–	–	–
SVM	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	59.31	51.69	57.64	55.50
Literature	Supervised approach proposed by Zhu et al. [112]																							70.28

Twitter2013

Approach	1%				5%				10%				20%				40%				100%			
	F1-pos	F1-neg	F1-neu	$\overline{F1}$	F1-pos	F1-neg	F1-neu	$\overline{F1}$	F1-pos	F1-neg	F1-neu	$\overline{F1}$	F1-pos	F1-neg	F1-neu	$\overline{F1}$	F1-pos	F1-neg	F1-neu	$\overline{F1}$	F1-pos	F1-neg	F1-neu	$\overline{F1}$
Self-training	40.97 ±6.18	26.10 ±6.88	61.27 ±2.59	38.03 ±4.81	60.63 ±2.06	**41.55 ±1.79**	67.37 ±0.87	**51.09 ±1.77**	63.85 ±1.01	**44.96 ±1.76**	**69.25 ±0.32**	**54.41 ±1.22**	**66.46 ±0.63**	**49.09 ±1.05**	**70.25 ±0.31**	**57.77 ±0.67**	**67.69 ±0.44**	**50.57 ±0.70**	**71.27 ±0.49**	**59.13 ±0.35**	–	–	–	–
Co-training	**66.80 ±1.47**	**35.76 ±13.13**	**67.51 ±2.08**	**51.28 ±7.14**	**65.27 ±3.17**	36.55 ±5.92	**68.87 ±0.47**	50.91 ±3.69	**64.59 ±1.88**	38.09 ±4.20	68.68 ±0.50	51.34 ±2.10	65.21 ±1.19	42.32 ±2.19	69.38 ±0.40	53.76 ±1.54	65.54 ±0.79	43.76 ±1.37	70.25 ±0.39	54.65 ±0.93	–	–	–	–
SVM	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	70.70	61.41	65.81	66.05
Literature	Supervised approach proposed by Miura et al. [59]																							72.12

Twitter2014

Approach	1%				5%				10%				20%				40%				100%			
	F1-pos	F1-neg	F1-neu	$\overline{F1}$	F1-pos	F1-neg	F1-neu	$\overline{F1}$	F1-pos	F1-neg	F1-neu	$\overline{F1}$	F1-pos	F1-neg	F1-neu	$\overline{F1}$	F1-pos	F1-neg	F1-neu	$\overline{F1}$	F1-pos	F1-neg	F1-neu	$\overline{F1}$
Self-training	50.22 ±8.16	27.92 ±5.83	54.04 ±2.24	39.07 ±5.99	59.51 ±2.13	**41.92 ±1.70**	59.41 ±0.75	50.71 ±1.56	62.72 ±1.30	**45.98 ±0.89**	61.35 ±0.54	**54.35 ±0.77**	**66.61 ±0.67**	**48.12 ±1.64**	63.29 ±0.44	**57.36 ±0.88**	**68.07 ±0.36**	**49.82 ±1.30**	**64.39 ±0.38**	**58.95 ±0.62**	–	–	–	–
Co-training	**67.93 ±2.26**	**38.69 ±14.00**	**61.40 ±1.44**	**53.31 ±7.65**	**66.41 ±2.95**	41.17 ±6.64	**62.99 ±0.67**	**53.79 ±4.13**	**66.14 ±2.31**	43.44 ±4.44	**63.04 ±0.61**	54.79 ±2.25	67.05 ±1.54	45.77 ±1.26	**63.78 ±0.62**	56.41 ±0.98	66.92 ±0.89	45.17 ±2.18	64.21 ±0.56	56.04 ±1.33	–	–	–	–
SVM	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	71.87	55.51	57.73	63.69
Literature	Supervised approach proposed by Miura et al. [59]																							70.96

Twitter Sarcasm 2014

Approach	1%				5%				10%				20%				40%				100%			
	F1-pos	F1-neg	F1-neu	$\overline{F1}$	F1-pos	F1-neg	F1-neu	$\overline{F1}$	F1-pos	F1-neg	F1-neu	$\overline{F1}$	F1-pos	F1-neg	F1-neu	$\overline{F1}$	F1-pos	F1-neg	F1-neu	$\overline{F1}$	F1-pos	F1-neg	F1-neu	$\overline{F1}$
Self-training	41.08 ±11.13	**16.97 ±10.68**	32.26 ±4.70	29.02 ±7.32	**49.76 ±4.87**	**19.54 ±6.46**	37.00 ±5.46	**34.65 ±4.23**	**49.90 ±1.53**	**19.90 ±4.93**	**37.62 ±1.11**	**34.90 ±2.26**	**54.18 ±2.09**	**22.44 ±2.45**	**42.66 ±1.50**	**38.31 ±1.91**	**54.54 ±2.17**	**24.76 ±2.36**	**44.83 ±3.39**	**39.65 ±1.30**	–	–	–	–
Co-training	**51.13 ±4.71**	10.50 ±4.25	**40.90 ±6.19**	**29.82 ±3.92**	48.81 ±5.58	10.29 ±4.06	**38.15 ±3.39**	29.55 ±3.48	47.82 ±6.45	9.34 ±4.82	36.49 ±1.86	28.58 ±3.49	45.60 ±4.08	10.79 ±3.07	37.04 ±1.69	28.20 ±3.06	43.42 ±4.12	13.47 ±3.02	36.90 ±1.38	28.44 ±2.48	–	–	–	–
SVM	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	57.15	25.00	50.00	41.08
Literature	Supervised approach proposed by Zhu et al. [112]																							58.16

in Table 3 for each test set. Note that, self-training in general provides better F-scores than co-training and is more stable in most cases as well.

Table 4: Number of higher F-scores and lower standard deviations (%) for self-training and co-training according to the results from Table 3.

Dataset	**Self-training**		**Co-training**	
	Higher F1s	*Lower Stds*	*Higher F1s*	*Lower Stds*
LiveJournal	**70**%	**65**%	30%	35%
SMS2013	**65**%	**50**%	55%	**50**%
Twitter2013	**65**%	**80**%	35%	20%
Twitter2014	**50**%	**80**%	**50**%	20%
Twitter Sarcasm 2014	**80**%	**60**%	20%	40%

4.4.3 Distant supervision based classification

We run experiments with an unsupervised approach known as tweet sentiment classification using distant supervision [27]. In this approach, the sentiment classes from the training set are replaced by positive and negative emoticons[13] and hashtags [61]. A tweet is considered as neutral when it does not contain emoticons or hashtags. Therefore, from distant supervision there is no human effort in the annotation of the data. In our experiments, new training sets were created with the same tweets, but their classes were based on the presence or absence of emoticons and sentiment hashtags.

Table 5 summarizes the results with tweet sentiment classification using distant supervision. All results are worse compared to self-training and co-training (even considering only 1% of labeled tweets). This occurs due to the small percentage of tweets in this data set with emotions and sentiment hashtags, since only 842 tweets had emoticons or sentiment hashtags (what represents 7.4% of the training set).

As mentioned in Section 2, the distant supervision based classification has been widely used. However, it requires large data sets to achieve satisfactory prediction power. For example, [84] collected 60M tweets over a two-month period and [94] collected 10 million tweets in April, 2013.

Table 5: Distant supervision results on the five test sets.

LiveJournal2014								
positive			negative			neutral		
precision	recall	F1	precision	recall	F1	precision	recall	F1
6.56	62.22	11.86	0.33	100.00	0.66	96.84	36.31	52.82
$\overline{F1}$: 6.26								
SMS2013								
positive			negative			neutral		
precision	recall	F1	precision	recall	F1	precision	recall	F1
5.49	60.00	10.06	0.51	40.00	1.00	98.51	58.20	73.17
$\overline{F1}$:5.53								
Twitter2013								
positive			negative			neutral		
precision	recall	F1	precision	recall	F1	precision	recall	F1
11.45	76.60	19.92	0.33	40.00	0.66	97.38	44.70	61.27
$\overline{F1}$: 10.29								
Twitter2014								
positive			negative			neutral		
precision	recall	F1	precision	recall	F1	precision	recall	F1
11.41	86.82	20.16	1.49	75.00	2.91	98.06	38.14	54.92
$\overline{F1}$:11.54								
Twitter Sarcasm 2014								
positive			negative			neutral		
precision	recall	F1	precision	recall	F1	precision	recall	F1
3.03	33.33	5.56	0.00	0.00	0.00	92.31	14.46	25.00
$\overline{F1}$:2.78								

[13] `http://en.wikipedia.org/wiki/List_of_emoticons`

5 Conclusions

We surveyed SSL approaches applied to tweet sentiment analysis. Tweet applications with semi-supervised settings are motivated by the fact that labeled tweets are typically expensive and difficult to obtain, whereas unlabeled tweets are generally widely available at low cost. Aiming to provide additional instructive guidelines for those interested in SSL-based tweet classification approaches, we also reported an experimental comparative analysis on real-world data from state-of-the-art algorithms. From this perspective, our study is helpful for new developments of human-machine systems for tweet sentiment analysis.

We empirically compared three SSL approaches namely: Self-training, Co-training, and Topic Modeling. In general, Co-training performed better without the presence of irony and sarcasm and with limited data (i.e., using at most 5% of the labeled data available). However, Self-training is the best choice when a significant amount of labeled tweets are available. In addition, the Self-training approach was observed to be more useful when irony and sarcasm are present. Considering samplings with 60% of the training set and a confidence threshold of 0.9, some learning progress has been observed for the Topic-based approach, but such a performance is still not compatible with those shown by the Self-training and Co-training algorithms.

As an emerging research topic, the use of SSL to address tweet sentiment analysis faces many challenges that motivate relevant future work such as:

1. When selecting a portion of available data for the initial training phase, some features may not make sense for the purpose of building classifiers. Therefore, it is important to evaluate the impact of the chosen features in a semi-supervised setting. In our experiments, the selected features were inspired in [61]; these authors ranked first in SemEval 2013 [63]. However, this feature set was defined based on the entire training set. It is more realistic to select features on the fly as an intrinsic part of the SSL. In addition, techniques for dimensionality reduction as principal component analysis (PCA), information gain, correlation-based feature selection (CFS) [13, 64] and feature hashing [18] are worth studying.

2. To increase the performance of SSL methods in applications where sarcasm and irony are present, it is interesting to study specific features, such as in [10, 30, 101, 22]. The feature set used in our experiments did not consider this particular scenario.

3. In [106], a topic model must be inferred from training data. The process of learning and updating the model in semi-supervised phase has a high computational cost, so this approach has shown to be ineffective in our experiments. We believe that combining classification and clustering for tweet sentiment analysis in a semi-supervised approach is promising, and should be explored with a topic model from the testing set instead of from the training set as (unrealistically) done in [106]. One idea is to capture the similarities among the tweets that are being classified, such

that the classifier can be refined from additional information provided by clusterers, as proposed in [17, 15].

4. Most SSL algorithms have been designed for binary classification. In tweet sentiment analysis these approaches have been extended to multi-class classification without any adaptation. Additional problems with extending semi-supervised binary classifiers to multi-class problems include imbalanced classification and different output scales of different binary classifiers. To adopt a binary SSL algorithm to problems with more than two classes, such as speech recognition and object recognition, multi-class problems are usually decomposed into a number of independent binary classification problems using techniques such as one-versus-the-rest, one-versus-one, and error-correcting output coding [24]. This type of solution has not yet been applied in tweet sentiment analysis and is a promising future work.

5. Given that the learning process of semi-supervised classifiers still depends on a small initial sample of labeled data, another interesting research topic involves studying sampling methods. [117] combines active and SSL in a Gaussian random field model, and indicates that the active learning scheme requires a much smaller number of queries to active high accuracy compared to random query selection. Recently, active learning has been applied in sentiment classification of movies and product reviews [20] with success, thereby suggesting that they can be applied to tweet sentiment analysis.

6. A dynamic and online tweet sentiment analysis is also an interesting research area that was not addressed in this paper. It has been studied for different applications [19, 25], with few studies for sentiment analysis [52, 41, 58, 6, 5], specially when a semi-supervised setting is considered [43, 54].

6 Acknowledgments

The authors would like to acknowledge Brazilian Research Agencies Capes (Proc. DS-7253238/D), CNPq (Proc. 303348/2013-5) and FAPESP (Proc. 2013/07375-0 and 2010/20830-0) for the financial support.

References

[1] Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. Sentiment analysis of twitter data. In *Proceedings of the Workshop on Languages in Social Media*, LSM '11, pages 30–38, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[2] Maria-Florina Balcan and Avrim Blum. A discriminative model for semi-supervised learning. *J. ACM*, 57(3):19:1–19:46, 2010.

[3] Wesley Baugh. bwbaugh : Hierarchical sentiment analysis with partial self-training. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 539–542, Atlanta, Georgia, USA, June 2013. Association for Computational Linguistics.

[4] Lee Becker, George Erhart, David Skiba, and Valentine Matula. Avaya: Sentiment analysis on twitter with self-training and polarity lexicon expansion. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 333–340, Atlanta, Georgia, USA, June 2013. Association for Computational Linguistics.

[5] Albert Bifet and Eibe Frank. Sentiment knowledge discovery in twitter streaming data. In *Proceedings of the 13th International Conference on Discovery Science*, DS'10, pages 1–15, Berlin, Heidelberg, 2010. Springer-Verlag.

[6] Albert Bifet, Geoffrey Holmes, and Bernhard Pfahringer. Moa-tweetreader: Real-time analysis in twitter streaming data. In *Proceedings of the 14th International Conference on Discovery Science*, DS'11, pages 46–60, Berlin, Heidelberg, 2011. Springer-Verlag.

[7] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003.

[8] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, COLT' 98, pages 92–100, New York, NY, USA, 1998. ACM.

[9] Pedro Henrique Calais Guerra, Adriano Veloso, Wagner Meira, Jr., and Virgílio Almeida. From bias to opinion: A transfer-learning approach to real-time sentiment analysis. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 150–158, New York, NY, USA, 2011. ACM.

[10] Paula Carvalho, Luís Sarmento, Mário J. Silva, and Eugénio de Oliveira. Clues for detecting irony in user-generated contents: Oh...!! it's "so easy" ;-). In *Proceedings of the 1st International CIKM Workshop on Topic-sentiment Analysis for Mass Opinion*, TSA '09, pages 53–56, New York, NY, USA, 2009. ACM.

[11] O. Chapelle, B. Schölkopf, and A. Zien. *Semi-Supervised Learning*. MIT Press, Cambridge, MA, 2006.

[12] Marc Cheong and Vincent C. Lee. A microblogging-based approach to terrorism informatics: Exploration and chronicling civilian sentiment and response to terrorism events via twitter. *Information Systems Frontiers*, 13(1):45–59, March 2011.

[13] Terence Tai-Leung Chong, Bingqing Cao, and Wing-Keung Wong. A new principal-component approach to measure the investor sentiment. In *IGEF Working Paper*, number 24, 2014.

[14] Sam Clark and Rich Wicentwoski. Swatcs: Combining simple classifiers with estimated accuracy. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 425–429, Atlanta, Georgia, USA, June 2013. Association for Computational Linguistics.

[15] Luiz F. S. Coletta, Eduardo R. Hruschka, Ayan Acharya, and Joydeep Ghosh. Using metaheuristics to optimize the combination of classifier and cluster ensembles. *Integrated Computer-Aided Engineering*, 22(3):229–242, 2015.

[16] Koby Crammer, Alex Kulesza, and Mark Dredze. Adaptive regularization of weight vectors. In *NIPS*, pages 414–422, 2009.

[17] Nadia F. F. da Silva, Luiz F. S. Coletta, Eduardo R. Hruschka, and Estevam R. Hruschka Jr. Using unsupervised information to improve semi-supervised tweet sentiment classification. *Information Sciences*, pages 1–18, 2016.

[18] Nadia F. F. da Silva, Eduardo R. Hruschka, and Estevam R. Hruschka Jr. Tweet sentiment analysis with classifier ensembles. *Decision Support Systems*, 66:170 – 179, 2014.

[19] N. Dahal, O. Abuomar, R. King, and V. Madani. Event stream processing for improved situational awareness in the smart grid. *Expert Systems with Applications*, 42(20):6853 – 6863, 2015.

[20] Sajib Dasgupta and Vincent Ng. Mine the easy, classify the hard: A semi-supervised approach to automatic sentiment classification. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2*, ACL '09, pages 701–709, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[21] Dmitry Davidov, Oren Tsur, and Ari Rappoport. Enhanced sentiment learning using twitter hashtags and smileys. In *Proceedings of the 23rd*

International Conference on Computational Linguistics: Posters, COLING '10, pages 241–249, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[22] Larissa A. de Freitas, Aline A. Vanin, Denise N. Hogetop, Marco N. Bochernitsan, and Renata Vieira. Pathways for irony detection in tweets. In *Proceedings of the 29th Annual ACM Symposium on Applied Computing*, SAC '14, pages 628–633, New York, NY, USA, 2014. ACM.

[23] Nicholas A. Diakopoulos and David A. Shamma. Characterizing debate performance via aggregated twitter sentiment. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 1195–1198, New York, NY, USA, 2010. ACM.

[24] Thomas G. Dietterich and Ghulum Bakiri. Solving multiclass learning problems via error-correcting output codes. *Journal of Artificial Intelligence Research*, 2:263–286, 1995.

[25] Pedro Domingos and Geoff Hulten. Mining high-speed data streams. In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '00, pages 71–80, New York, NY, USA, 2000. ACM.

[26] Ronen Feldman. Techniques and applications for sentiment analysis. *Commun. ACM*, 56(4):82–89, April 2013.

[27] Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. *Unpublished Manuscript, Stanford University*, pages 1–6, 2009.

[28] A. B. Goldberg. *New directions in semi-supervised learning*. PhD thesis, University of Wisconsin - Madison, 2010.

[29] Andrew B. Goldberg and Xiaojin Zhu. Seeing Stars when there aren't many Stars: Graph-based Semi-supervised Learning for Sentiment Categorization. In *Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing*, TextGraphs-1, pages 45–52, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

[30] Roberto González-Ibáñez, Smaranda Muresan, and Nina Wacholder. Identifying sarcasm in twitter: A closer look. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, HLT '11, pages 581–586, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[31] Geoffrey J. Gordon, David Blei Dunson, and Miroslav Dudík, editors. *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2011, Fort Lauderdale, USA, April 11-13, 2011*, volume 15 of *JMLR Proceedings*. JMLR.org, 2011.

[32] Yoav Haimovitch, Koby Crammer, and Shie Mannor. More is better: Large scale partially-supervised sentiment classification - appendix. *CoRR*, abs/1209.6329, 2012.

[33] Ammar Hassan, Ahmed Abbasi, and Daniel Zeng. Twitter sentiment analysis: A bootstrap ensemble framework. In *SocialCom*, pages 357–364. IEEE, 2013.

[34] Yulan He, Chenghua Lin, Wei Gao, and Kam-Fai Wong. Tracking sentiment and topic dynamics from social media. In *ICWSM*, 2012.

[35] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '04, pages 168–177, New York, NY, USA, 2004. ACM.

[36] Xia Hu, Lei Tang, Jiliang Tang, and Huan Liu. Exploiting social relations for sentiment analysis in microblogging. In *Proceedings of the sixth ACM international conference on Web search and data mining*, 2013.

[37] Bernard J. Jansen, Mimi Zhang, Kate Sobel, and Abdur Chowdury. Twitter power: Tweets as electronic word of mouth. *J. Am. Soc. Inf. Sci. Technol.*, 60(11):2169–2188, nov 2009.

[38] Yohan Jo and Alice H. Oh. Aspect and sentiment unification model for online review analysis. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM '11, pages 815–824, New York, NY, USA, 2011. ACM.

[39] Thorsten Joachims. Advances in kernel methods. chapter Making Large-scale Support Vector Machine Learning Practical, pages 169–184. MIT Press, Cambridge, MA, USA, 1999.

[40] Christopher Johnson, Parul Shukla, and Shilpa Shukla. On classifying the political sentiment of tweets, 2011.

[41] Hwi-Gang Kim, Seongjoo Lee, and Sunghyon Kyeong. Discovering hot topics using twitter streaming data: Social topic detection and geographic clustering. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '13, pages 1215–1220, New York, NY, USA, 2013. ACM.

[42] Soo-Min Kim and Eduard Hovy. Determining the sentiment of opinions. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics.

[43] Milosz R. Kmieciak and Jerzy Stefanowski. Handling sudden concept drift in enron messages data stream. *Control and Cybernetics*, 40(3):667–695, 2011.

[44] P. F. Lazarsfeld and R. K. Merton. Friendship as a social process: A substantive and methodological analysis. In M. Berger, T. Abel, and C. Page, editors, *Freedom and Control in Modern Society*, pages 18–66. Van Nostrand, New York, 1954.

[45] Shoushan Li, Zhongqing Wang, Guodong Zhou, and Sophia Yat Mei Lee. Semi-supervised learning for imbalanced sentiment classification. In *IJCAI*, pages 1826–1831, 2011.

[46] Tao Li, Vikas Sindhwani, Chris H. Q. Ding, and Yi Zhang. Bridging domains with words: Opinion analysis with matrix tri-factorizations. In *SDM*, pages 293–302, 2010.

[47] Jimmy Lin and Alek Kolcz. Large-scale machine learning at twitter. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, SIGMOD '12, pages 793–804, New York, NY, USA, 2012. ACM.

[48] Bing Liu. *Sentiment Analysis and Opinion Mining*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2012.

[49] Shenghua Liu, Fuxin Li, Fangtao Li, Xueqi Cheng, and Huawei Shen. Adaptive co-training svm for sentiment classification on tweets. In *Proceedings of the 22Nd ACM International Conference on Conference on Information & Knowledge Management*, CIKM '13, pages 2079–2088, New York, NY, USA, 2013. ACM.

[50] Shenghua Liu, Wenjun Zhu, Ning Xu, Fangtao Li, Xue-qi Cheng, Yue Liu, and Yuanzhuo Wang. Co-training and visualizing sentiment evolvement for tweet events. In *Proceedings of the 22Nd International Conference on World Wide Web Companion*, WWW '13 Companion, pages 105–106, Republic and Canton of Geneva, Switzerland, 2013. International World Wide Web Conferences Steering Committee.

[51] Zhiguang Liu, Xishuang Dong, Yi Guan, and Jinfeng Yang. Reserved self-training: A semi-supervised sentiment classification method for chinese microblogs. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 455–462, Nagoya, Japan, October 2013. Asian Federation of Natural Language Processing.

[52] Roberto Lourenco Jr., Adriano Veloso, Adriano Pereira, Wagner Meira Jr., Renato Ferreira, and Srinivasan Parthasarathy. Economically-efficient sentiment stream analysis. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '14, pages 637–646, New York, NY, USA, 2014. ACM.

[53] Yue Lu, Panayiotis Tsaparas, Alexandros Ntoulas, and Livia Polanyi. Exploiting social context for review quality prediction. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 691–700, New York, NY, USA, 2010. ACM.

[54] Mohammad Masud, Jing Gao, Latifur Khan, Jiawei Han, and Xiaohu Li. A practical approach to classify evolving data streams: Training with limited amount of labeled data. *Proc. 2008 Int. Conf. on Data Mining (ICDM'08), Pisa, Italy, Dec. 2008*, December 2008.

[55] Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27(1):415–444, 2001.

[56] Walaa Medhat, Ahmed Hassan, and Hoda Korashy. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4):1093 – 1113, 2014.

[57] Qiaozhu Mei, Xu Ling, Matthew Wondra, Hang Su, and ChengXiang Zhai. Topic sentiment mixture: Modeling facets and opinions in weblogs. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pages 171–180, New York, NY, USA, 2007. ACM.

[58] Yelena Mejova and Padmini Srinivasan. Political speech in social media streams: Youtube comments and twitter posts. In *Proceedings of the 4th Annual ACM Web Science Conference*, WebSci '12, pages 205–208, New York, NY, USA, 2012. ACM.

[59] Yasuhide Miura, Shigeyuki Sakaki, Keigo Hattori, and Tomoko Ohkuma. Teamx: A sentiment analyzer with enhanced lexicon mapping and weighting scheme for unbalanced data. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 628–632, Dublin, Ireland, August 2014. Association for Computational Linguistics and Dublin City University.

[60] Saif Mohammad. #emotional tweets. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 246–255, Montréal, Canada, 7-8 June 2012. Association for Computational Linguistics.

[61] Saif Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. In *Proceedings of the seventh international workshop on Semantic Evaluation Exercises (SemEval-2013)*, Atlanta, Georgia, USA, June 2013.

[62] AndréS Montoyo, Patricio MartíNez-Barco, and Alexandra Balahur. Subjectivity and sentiment analysis: An overview of the current state of the area and envisaged developments. *Decis. Support Syst.*, 53(4):675–679, November 2012.

[63] Preslav Nakov, Sara Rosenthal, Zornitsa Kozareva, Veselin Stoyanov, Alan Ritter, and Theresa Wilson. Semeval-2013 task 2: Sentiment analysis in twitter. In *Second Joint Conference on Lexical and Computational*

*Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 312–320, Atlanta, Georgia, USA, June 2013. Association for Computational Linguistics.

[64] Tim O'Keefe and Irena Koprinska. Feature Selection andWeighting Methods in Sentiment Analysis. In *Proceedings of 14th Australasian Document Computing Symposium*, December 2009.

[65] Olutobi Owoputi, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A. Smith. Improved part-of-speech tagging for online conversational text with word clusters. In *In Proceedings of NAACL*, 2013.

[66] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Trans. on Knowl. and Data Eng.*, 22(10):1345–1359, October 2010.

[67] Bo Pang and Lillian Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of ACL*, pages 115–124, 2005.

[68] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of EMNLP*, pages 79–86, 2002.

[69] FedericoAlberto Pozzi, Daniele Maccagnola, Elisabetta Fersini, and Enza Messina. Enhance user-level sentiment analysis on microblogs with approval relations. In Matteo Baldoni, Cristina Baroglio, Guido Boella, and Roberto Micalizio, editors, *AI*IA 2013: Advances in Artificial Intelligence*, volume 8249 of *Lecture Notes in Computer Science*, pages 133–144. Springer International Publishing, 2013.

[70] Ashequl Qadir and Ellen Riloff. Bootstrapped learning of emotion hashtags #hashtags4you. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 2–11, Atlanta, Georgia, June 2013. Association for Computational Linguistics.

[71] Guang Qiu, Xiaofei He, Feng Zhang, Yuan Shi, Jiajun Bu, and Chun Chen. Dasa: Dissatisfaction-oriented advertising based on sentiment analysis. *Expert Systems with Applications*, 37(9):6182 – 6191, 2010.

[72] Likun Qiu, Weishi Zhang, Changjian Hu, and Kai Zhao. Selc: A self-supervised model for sentiment classification. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pages 929–936, New York, NY, USA, 2009. ACM.

[73] Delip Rao and David Yarowsky. Ranking and semi-supervised classification on large scale graphs using map-reduce. In *Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing*,

TextGraphs-4, pages 58–65, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[74] Jonathon Read. Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In *Proceedings of the ACL Student Research Workshop*, ACLstudent '05, pages 43–48, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[75] Jonathon Read and John Carroll. Weakly supervised techniques for domain-independent sentiment classification. In *Proceedings of the 1st International CIKM Workshop on Topic-sentiment Analysis for Mass Opinion*, TSA '09, pages 45–52, New York, NY, USA, 2009. ACM.

[76] Yong Ren, Nobuhiro Kaji, Naoki Yoshinaga, and Masaru Kitsuregawa. Sentiment classification in under-resourced languages using graph-based semi-supervised learning methods. *IEICE Transactions*, 97-D(4):790–797, 2014.

[77] Carlos Rodrıguez-Penagos, Jordi Atserias, Joan Codina-Filba, David Garcıa-Narbona, Jens Grivolla, Patrik Lambert, and Roser Saurı. Fbm: Combining lexicon-based ml and heuristics for social media polarities. In *Proceedings of SemEval-2013 – International Workshop on Semantic Evaluation Co-located with *Sem and NAACL*, Atlanta, Georgia, 2013. Url date at 2013-10-10.

[78] Sara Rosenthal, Preslav Nakov, Alan Ritter, and Veselin Stoyanov. SemEval-2014 Task 9: Sentiment Analysis in Twitter. In Preslav Nakov and Torsten Zesch, editors, *Proceedings of the 8th International Workshop on Semantic Evaluation*, SemEval 14, Dublin, Ireland, 2014.

[79] Hassan Saif, Yulan He, and Harith Alani. Semantic sentiment analysis of twitter. In *Proceedings of the 11th international conference on The Semantic Web - Volume Part I*, ISWC'12, pages 508–524, Berlin, Heidelberg, 2012. Springer-Verlag.

[80] Hassan Saif, Yulan He, Miriam Fernandez, and Harith Alani. Contextual semantics for sentiment analysis of twitter. *Information Processing and Management*, (0):–, 2015.

[81] Ted Sandler, John Blitzer, Partha Pratim Talukdar, and Lyle H. Ungar. Regularized learning with networks of features. In Daphne Koller, Dale Schuurmans, Yoshua Bengio, and Léon Bottou, editors, *NIPS*, pages 1401–1408. Curran Associates, Inc., 2008.

[82] III Scudder, H. Probability of error of some adaptive pattern-recognition machines. *Information Theory, IEEE Transactions on*, 11(3):363–371, Jul 1965.

[83] Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM Comput. Surv.*, 34(1):1–47, March 2002.

[84] Aliaksei Severyn and Alessandro Moschitti. Twitter sentiment analysis with deep convolutional neural networks. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, pages 959–962, New York, NY, USA, 2015. ACM.

[85] Jianfeng Si, Arjun Mukherjee, Bing Liu, Qing Li, Huayi Li, and Xiaotie Deng. Exploiting topic based twitter sentiment for stock prediction. In *ACL (2)*, pages 24–29, 2013.

[86] Nádia Silva, Estevam Hruschka, and Eduardo Hruschka. Biocom usp: Tweet sentiment analysis with adaptive boosting ensemble. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 123–128, Dublin, Ireland, August 2014. Association for Computational Linguistics and Dublin City University.

[87] Vikas Sindhwani and Prem Melville. Document-word co-regularization for semi-supervised sentiment analysis. In *ICDM*, pages 1025–1030, 2008.

[88] Jared Suttles and Nancy Ide. Distant supervision for emotion classification with discrete binary values. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 7817 of *Lecture Notes in Computer Science*, pages 121–136. Springer Berlin Heidelberg, 2013.

[89] Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. Lexicon-based methods for sentiment analysis. *Comput. Linguist.*, 37(2):267–307, June 2011.

[90] Chenhao Tan, Lillian Lee, Jie Tang, Long Jiang, Ming Zhou, and Ping Li. User-level sentiment analysis incorporating social networks. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 1397–1405, New York, NY, USA, 2011. ACM.

[91] Duyu Tang, Bing Qin, and Ting Liu. Deep learning for sentiment analysis: successful approaches and future challenges. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 5(6):292–303, 2015.

[92] Duyu Tang, Furu Wei, Bing Qin, Ting Liu, and Ming Zhou. Coooolll: A deep learning system for twitter sentiment classification. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 208–212, Dublin, Ireland, August 2014. Association for Computational Linguistics and Dublin City University.

[93] Duyu Tang, Furu Wei, Bing Qin, Ming Zhou, and Ting Liu. Building large-scale twitter-specific sentiment lexicon : A representation learning approach. In *COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, August 23-29, 2014, Dublin, Ireland*, pages 172–182, 2014.

[94] Duyu Tang, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, and Bing Qin. Learning sentiment-specific word embedding for twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1: Long Papers*, pages 1555–1565, 2014.

[95] Harsh Thakkar and Dhiren Patel. Approaches for sentiment analysis on twitter: A state-of-art study. In *International Network for Social Network Analysis conference (INSNA)*, Xi'an, China, 2013.

[96] Mike Thelwall. Emotion homophily in social network site messages. *First Monday*, 15(4), 2010.

[97] Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. Sentiment in short strength detection informal text. *J. Am. Soc. Inf. Sci. Technol.*, 61(12):2544–2558, December 2010.

[98] Mikalai Tsytsarau and Themis Palpanas. Survey on mining subjective data on the web. *Data Min. Knowl. Discov.*, 24(3):478–514, May 2012.

[99] Peter D. Turney. Mining the web for synonyms: Pmi-ir versus lsa on toefl. In *Proceedings of the 12th European Conference on Machine Learning*, EMCL '01, pages 491–502, London, UK, UK, 2001. Springer-Verlag.

[100] Peter D. Turney. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL 02, pages 417–424, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[101] Aline A. Vanin, Larissa A. Freitas, Renata Vieira, and Marco Bochernitsan. Some clues on irony detection in tweets. In *Proceedings of the 22Nd International Conference on World Wide Web Companion*, WWW '13 Companion, pages 635–636, Republic and Canton of Geneva, Switzerland, 2013. International World Wide Web Conferences Steering Committee.

[102] G. Vinodhini and R. M. Chandrasekaran. Sentiment Analysis and Opinion Mining: A Survey. *International Journal*, 2(6), 2012.

[103] Xiaojun Wan. Co-training for cross-lingual sentiment classification. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - Volume 1*, ACL '09, pages 235–243, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[104] Changbo Wang, Zhao Xiao, Yuhua Liu, Yanru Xu, Aoying Zhou, and Kang Zhang. Sentiview: Sentiment analysis and visualization for internet popular topics. *IEEE T. Human-Machine Systems*, 43(6):620–630, 2013.

[105] Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan, and Amit P. Sheth. Harnessing twitter "big data" for automatic emotion identification. In *SocialCom/PASSAT*, pages 587–592. IEEE, 2012.

[106] Bing Xiang and Liang Zhou. Improving twitter sentiment analysis with topic-based mixture modeling and semi-supervised training. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 434–439, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

[107] Jun-Ming Xu, Kwang-Sung Jun, Xiaojin Zhu, and Amy Bellmore. Learning from bullying traces in social media. In *HLT-NAACL*, pages 656–666. The Association for Computational Linguistics, 2012.

[108] Ning Yu. Exploring co-training strategies for opinion detection. *Journal of the Association for Information Science and Technology*, pages n/a–n/a, 2014.

[109] Taras Zagibalov and John Carroll. Unsupervised classification of sentiment and objectivity in Chinese text. In *Proceedings of the Third International Joint Conference on Natural Language Processing (IJCNLP)*, 2008.

[110] Lei Zhang, Riddhiman Ghosh, Mohamed Dekhil, Meichun Hsu, and Bing Liu. Combining lexicon-based and learning-based methods for twitter sentiment analysis. HP Laboratories, 2011. Technical Report.

[111] Jiang Zhao, Man Lan, and Tiantian Zhu. Ecnu: Expression- and message-level sentiment orientation classification in twitter using multiple effective features. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 259–264, Dublin, Ireland, August 2014. Association for Computational Linguistics and Dublin City University.

[112] Xiaodan Zhu, Svetlana Kiritchenko, and Saif Mohammad. Nrc-canada-2014: Recent improvements in the sentiment analysis of tweets. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 443–447, Dublin, Ireland, August 2014. Association for Computational Linguistics and Dublin City University.

[113] Xiaojin Zhu. Semi-supervised learning literature survey. 2005. Technical Report 1530, Computer Sciences, University of Wisconsin-Madison.

[114] Xiaojin Zhu and Andrew B Goldberg. Kernel regression with order preferences. *Proceedings of the National Conference on Artificial Intelligence*, 22(1):681, 2007.

[115] Xiaojin Zhu and Andrew B. Goldberg. *Introduction to Semi-Supervised Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers, 2009.

[116] Xiaojin Zhu, Andrew B. Goldberg, Ronald Brachman, and Thomas Dietterich. *Introduction to Semi-Supervised Learning.* Morgan and Claypool Publishers, 2009.

[117] Xiaojin Zhu, John Lafferty, and Zoubin Ghahramani. Combining active learning and semi-supervised learning using gaussian fields and harmonic functions. In *ICML 2003 workshop on The Continuum from Labeled to Unlabeled Data in Machine Learning and Data Mining*, pages 58–65, 2003.