Current State of Text Sentiment Analysis from Opinion to Emotion Mining

ALI YADOLLAHI, AMENEH GHOLIPOUR SHAHRAKI, and OSMAR R. ZAIANE,
University of Alberta

Sentiment analysis from text consists of extracting information about opinions, sentiments, and even emotions conveyed by writers towards topics of interest. It is often equated to opinion mining, but it should also encompass emotion mining. Opinion mining involves the use of natural language processing and machine learning to determine the attitude of a writer towards a subject. Emotion mining is also using similar technologies but is concerned with detecting and classifying writers emotions toward events or topics. Textual emotion-mining methods have various applications, including gaining information about customer satisfaction, helping in selecting teaching materials in e-learning, recommending products based on users emotions, and even predicting mental-health disorders. In surveys on sentiment analysis, which are often old or incomplete, the strong link between opinion mining and emotion mining is understated. This motivates the need for a different and new perspective on the literature on sentiment analysis, with a focus on emotion mining. We present the state-of-the-art methods and propose the following contributions: (1) a taxonomy of sentiment analysis; (2) a survey on polarity classification methods and resources, especially those related to emotion mining; (3) a complete survey on emotion theories and emotion-mining research; and (4) some useful resources, including lexicons and datasets.

1. INTRODUCTION

"Sentiment analysis," one of the fields in "affective computing," refers to all the areas of detecting, analyzing, and evaluating humans' state of mind towards different events, issues, services, or any other interest. More precisely, this field aims to mine opinions, sentiments, and emotions based on observations of people's actions that can be captured using their writings, facial expressions, speech, music, movements, and so on. Analysis of sentiments from each of these media is a specific field of study. Here we focus only on text sentiment analysis. For further information regarding other types of sentiment analysis, one can refer to Yang and Chen [2012], El Ayadi et al. [2011], Zeng et al. [2009], Kleinsmith and Bianchi-Berthouze [2013], and D'mello and Kory [2015].

Authors' addresses: A. Yadollahi, A. Gholipour Shahraki, and O. R. Zaiane, Computing Science Department, University of Alberta, 4-43 Athabasca Hall, Edmonton, Alberta, Canada T6G 2E8; emails: {yadollah, ameneh, zaiane}@ualberta.ca.
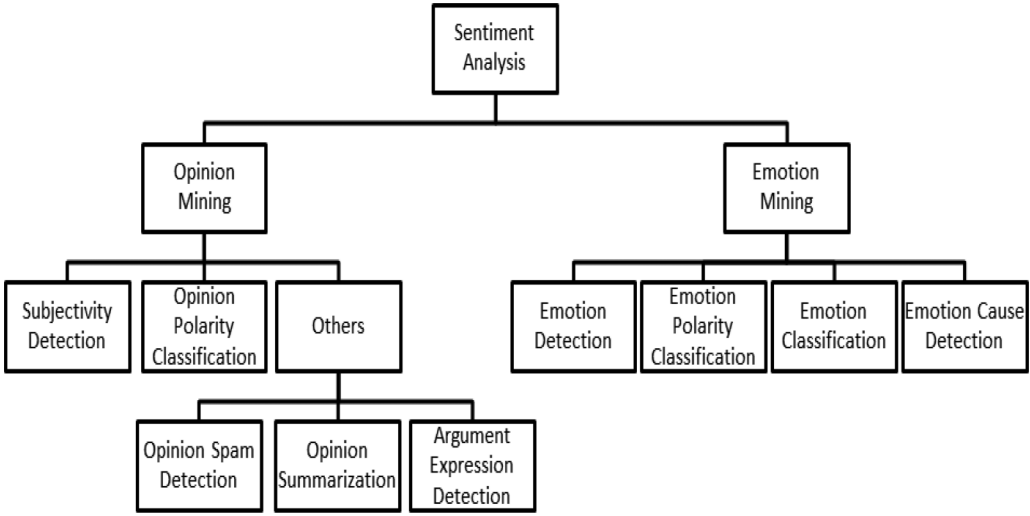
25

Fig. 1.   Taxonomy of sentiment analysis tasks.

Text sentiment analysis has been an attractive topic of study since the mid-1990s; however, there barely exists a systematic organization of tasks under this area and people use different terms to refer to different tasks. For example, *sentiment analysis*, *opinion mining*, and *polarity classification*, which we will define below, are used to address the same concept, while this is not sound either lexically or semantically. This is why having a clear definition of terms and a logical taxonomy of sentiment analysis work is one of our concerns.

According to the definition in the Merriam-Webster dictionary, sentiment is an attitude, thought, or judgment prompted by a feeling. In other words, sentiment is an opinion or idea colored by an emotion. Therefore, analyzing the sentiment of a unit of text can encompass investigating both the opinion and the emotion behind that unit.

It is easy to confuse opinion and emotion, since they have a strong correlation. For instance, in many situations emotion motivates a person to judge an entity and build opinions about it. Additionally, opinion of a person can cause emotions in others. However, a text unit can indicate contradicting opinions and emotions. For instance, the sentence "My family thinks it's a good decision to continue my education overseas, though they feel sad to miss me" represents a positive opinion and a negative emotion toward the same topic.

Based on the aforementioned reasons, we categorize the field of sentiment analysis into two parts: (1) opinion mining, dealing with the expression of opinions, and (2) emotion mining, concerned with the articulation of emotions. Opinion mining is more concerned with the concept of opinions expressed in texts that can be *positive*, *negative*, or *neutral*, while emotion mining is the study of emotions (e.g., *joy*, *sadness*) reflected in a piece of text. Hence, to have a sound terminology of problems, we should discriminate them. Figure 1 shows the categorization of sentiment analysis to these two tasks and the subtasks of each. These subtasks are defined as follows.

Opinion-mining tasks:

—**Subjectivity Detection:** The task of detecting if a text is objective or subjective. Objective texts carry some factual information, for example, "The sky is blue," while subjective texts express somebody's personal views or opinions, for example, "I like the color blue" [Liu 2011].

—**Opinion Polarity Classification:** The task of determining whether the text expresses *positive* or *negative* (or sometimes *neutral*) opinion. As mentioned above, "sentiment analysis" and "opinion mining" are used as synonyms of "polarity classification," which is restrictive. Section 2 of this article discusses many of the previous works corresponding to this subtask.

—**Opinion Spam Detection:** The task of detecting fake opinions in favor of or against a product or service that malicious users intentionally write to make their target popular or unpopular. The work of Jindal and Liu [2008] is one of the first attempts with promising results in this area of study.

—**Opinion Summarization:** The task of summarizing a large bunch of opinions toward a topic, encompassing different perspectives, aspects, and polarities. This is important specifically when someone wants to make a decision, because a single opinion cannot be trustworthy. The work of Hu and Liu [2004] is an example of opinion summarization on product reviews.

—**Argument Expression Detection:** The task of identifying argumentative structures and the relationship between different arguments within a document, such as one being opposed to the other. The work of Lin et al. [2006] is one of the interesting previous works for one to read.

Emotion-mining tasks:

—**Emotion Detection:** The task of detecting if a text conveys any type of emotion or not. This is similar to subjectivity detection for opinions and is addressed in Gupta et al. [2013].

—**Emotion Polarity Classification:** The task of determining the polarity of the existing emotion in a text, assuming that it has some. This is similar to opinion polarity classification. Examples of this study can be found in Alm et al. [2005] and Hancock et al. [2007].

—**Emotion Classification:** The task of fine-grained classification of existing emotion in a text into one (or more) of a set of defined emotions. Most of the literature that we elaborate on later in this article falls into this category.

—**Emotion Cause Detection:** The task of mining factors for eliciting some kinds of emotions, as in the early work by Lee et al. [2010] and a later work by Gao et al. [2015b].

As can be inferred from the definitions, we discriminate the words "detection" and "classification." The answer to a detection problem (of an opinion or emotion) is yes or no, meaning that there exists any opinion or emotion in the text or not. However, the answer to a classification problem is the exact type of opinion (*positive*, *negative*) or emotion (*joy*, *sadness*, etc.) of the target text.

Besides, Figure 1 shows the discrimination among the terms *sentiment analysis*, *opinion mining*, and *polarity classification*. In the literature, all these terms are used to refer to the problem of opinion polarity classification; however, we see that opinion polarity classification is a subtask of opinion mining, where opinion mining, in turn, is a subtask of sentiment analysis. In this article, we use each term for its exact and specific task and differentiate among them.

1.1. Motivation

As Figure 1 shows, there is a rich body of research on opinion mining, and many focused and specialized areas are investigated, while emotion mining from text is still in its infancy and still has a long way to proceed. Emotion mining is an interesting topic in many disciplines such as neuroscience, cognitive sciences, and psychology. Only recently has it attracted attention in computer science. Developing systems that

can detect emotions from text has many potential applications. In customer care services, emotion mining can help marketers gain information about how much satisfied their customers are and what aspects of their service should be improved or revised to consequently make a strong relationship with their end users [Gupta et al. 2013]. Users' emotions can additionally be used for sale predictions of a particular product. In e-learning applications, the Intelligent Tutoring System can decide on teaching materials, based on user's feelings and mental state. In Human Computer Interaction, the computer can monitor user's emotions to suggest suitable music or movies [Voeffray 2011]. Having the technology of identifying emotions enables new textual access approaches such as allowing users to filter results of a search by emotion. In addition, output of an emotion-mining system can serve as input to other systems. For instance, Rangel and Rosso [2016] use the emotions detected in the text for author profiling, specifically identifying the writer's age and gender. Last but not least, psychologists can infer patients' emotions and predict their state of mind accordingly. On a longer period of time, they are able to detect if a patient is facing depression or stress [De Choudhury et al. 2013] or even thinks about committing suicide, which is extremely useful, since he/she can be referred to counseling services [Luyckx et al. 2012].

On the other hand, with the explosive growth of web 2.0 technology, different media are available for people to express themselves and their feelings. This has added another aspect to the area. There is research on detecting emotions from text, facial expressions, images, speeches, paintings, songs, and other sorts of media [Busso et al. 2004; Wieczorkowska et al. 2006]. Among all, facial expressions and voice recorded speeches contain the most dominant clues and have widely been studied. There are also studies on combination of different types of information such as features from text and image including the work of Zhang et al. [2015]. Here we focus only on text and therefore cannot take advantage of the information conveyed via facial or audio channels. Personal notes, emails, news headlines, blogs, tales, novels, and chat messages are some types of text that can convey emotions. Particularly, popular social networking websites such as Twitter, Facebook, and MySpace are appropriate places to share one's feelings easily and widely.

There exist some comprehensive surveys on sentiment analysis by Pang and Lee [2008] and Liu [2012], where the latter was expanded in Liu [2015]. While methods and techniques discussed in these articles can be applied to the field of emotion mining as well, none of them have specific coverage of this task. There are also some surveys focusing on emotion mining, such as the works by Kao et al. [2009] and Jain and Kulkarni [2014], but they are rather incomplete. In addition, most of the works on emotion mining do not consider the strong link between emotion and opinion mining. In fact, many of the methods and techniques used in opinion mining can also be applied to emotion-mining problems. These facts motivate us to cover the state-of-the-art methods and resources developed for this popular task by taking a sentiment-analysis-oriented perspective to be a complementary to existing sentiment analysis surveys.

In addition, as shown in Figure 1, polarity classification can be applied to both opinion and emotion; however, in the literature it is almost always referring to opinion polarity classification. For instance, Pang and Lee [2008] mention: "The binary classification task of labelling an opinionated document as expressing either an overall positive or an overall negative opinion is called sentiment polarity classification or polarity classification." Nevertheless, proposed techniques and methods are useful for emotion polarity classification as well for two reasons: (1) opinion and emotion are semantically related concepts. Generally, having an opinion towards an entity can cause the person to feel an emotion in the same direction (*positive* or *negative*), and (2) these techniques often do not have any opinion-specific characteristic, and, hence, they can directly be

applied to emotionally labeled problems, too. Considering this inference, we believe it is worth reviewing polarity classification methods before entering emotion research.

This article is organized as follows: In Section 2, key elements of the polarity classification task are explained, and those works in this area that can be useful for the emotion-mining task are reviewed. In Section 3, a set of important resources, including lexicons and datasets that researchers need for a polarity classification task, are introduced. Reviewing emotion theories in order to gain knowledge about basic emotions is done is Section 4. A thorough survey on emotion-related research is given in Section 5. Section 6 is dedicated to introducing useful resources specific to emotion-mining work, and, finally, Section 7 summarizes and concludes the discussion.

2. POLARITY CLASSIFICATION METHODOLOGIES

Polarity classification is the task of classifying the opinion of a given text as falling under one of two opposing sentiment polarities, the most famous of which is "like" vs. "dislike" [Pang and Lee 2008]. Although much of the work in this area has been done on products and services reviews, which mostly hold *positive* or *negative* opinions, there are other problems where "like" or "dislike" are interpreted as other concepts such as different political views [Pang and Lee 2008]. As stated in Section 1, different media can be used to express opinions, among which we only focus on text. For more information about other types of polarity classification, one can refer to Morency et al. [2011].

Automatic classification of polarity can be categorized with respect to various perspectives. In terms of granularity, it can be done on a document, sentence, or aspect level.

—**Document level:** In this category, the whole document, whether short or long, is the atomic unit of input to the problem, and the polarity of the whole document is the essence of the study. Document-level polarity classification concerns most of the body of the work for this area and is considered the simplest sentiment analysis task in the research community [Liu 2015]. At the same time, it is widely demanding, since most of the online data includes documents such as reviews, blog posts, and comments. Document-level polarity classification is an essential requirement for studies such as social and psychological studies in social networks [Ortigosa et al. 2014; Gao et al. 2015a], consumer satisfaction [Kang and Park 2014], and analyzing patients in medical settings [Denecke and Deng 2015].

—**Sentence level:** The objective of this group of studies is to determine the polarity of a sentence. As noted in Neviarouskaya et al. [2007], a challenge at this level is the influence of the surrounding context on the sentence. For example, depending on what context it is used, the sentence "*I can't really describe this product better than this*" can be both *positive* and *negative*. Polarity classification of tweets, which has been extensively studied in the recent years, is the most interesting application of sentence-level polarity classification.

—**Aspect level:** This category, also known as feature-based opinion mining, encompasses the study of discovering opinion polarities about a specific aspect of a product or service. For instance, opinions on restaurants can be about two aspects of quality, namely the food and the cleanliness of the restaurant. This category of works is highly useful for business owners and politicians to gain insights about aggregations of people's opinions regarding various features of their product and services, where document- or sentence-level classifications do not suffice.

Extraction of aspects from text and polarity classification of the extracted aspects are the two major components of aspect-level polarity classification. The work of Hu and Liu [2004] is one of the earliest in this field. Further attempts mostly focused

on enhancing only one of these components. For instance, one of the most important group of works in this category is devoted to utilizing topic modeling in aspect extraction such as the work of Lin and He [2009], Jo and Oh [2011], Mukherjee and Liu [2012], and Wang et al. [2016].

With respect to the nature of the data, there are two important modes of the problem. Some datasets benefit from being annotated by a human, while there are many unlabeled datasets of reviews and posts. Methods working with labeled data often show better results; nevertheless, they require manual labeling, which one might be unable to afford. In the following two subsections, we discuss previous methods on annotated and unannotated text data, respectively.

2.1. Works on Annotated Data

The algorithms that deal with labeled data are called "supervised methods." Supervised methods apply some machine-learning algorithms on a set of training data to be able to predict the label of unseen test data. They need an annotated dataset of texts for the task of training, which creates a model to discriminate between polarities.

In order to apply machine-learning methods, one should represent the text by means of descriptive features. After that, some techniques should be used to train a polarity classifier. Most solutions introduced in the literature are general-purpose machine-learning techniques, while some of them are sentiment specific. Sebastiani [2002] was the first to apply general text categorization algorithms on the field of sentiment detection. Later, Pang et al. [2002] compared performance of Support Vector Machine (SVM) and Naïve Bayes against each other for movie reviews.

Representation learning methods have shown promising classification results in various applications, one of which is the polarity classification. Socher et al. [2013] utilize deep learning to train a Treebank sentiment classifier, Tang et al. [2014a] develop a deep learning Twitter sentiment model, dos Santos and Gatti [2014] apply Deep Convolutional Neural Networks on classifying short text, Tang et al. [2014b] develop neural networks to find continuous word representation along with the sentiment of the word, and Tang [2015] attempts to encapsulate features of a document using cascaded constitutes and to learn sentiment of documents. All these works attempt to find a representation of the polarity by applying various layers of hidden nodes among which the first layer consists of the raw features of the text.

A fairly large part of the literature is dedicated to finding out the usefulness of many features and techniques in learning. The most common types of those features, which have been also applied in other areas of text mining, are as follows.

2.1.1. Presence-Based and Frequency-Based Features. The most common way to describe a piece of text is by using a binary vector in which each element corresponds to one term from a dictionary. The element at index i in the vector is set to 1 if the term i is present in the text and is 0 otherwise. Likewise, one may describe the text as a vector representing the number of times individual terms have been repeated. The former is called the presence-based and the latter is named the frequency-based type of feature. Although term frequency is a popular feature in information retrieval, Pang et al. [2002] obtain better performance when using presence-based features.

2.1.2. Unigram and N-Gram Features. A unigram refers to one single word in a text and an n-gram represents a group of adjacent words in a sentence, preserving the order. Although n-grams have more information than unigram features, concerning the position of words in the sentence and being used as a group, them being more effective in increasing the performance is a matter of some debate. For instance, Pang et al. [2002] report that unigrams are more effective than n-grams; however, some

other research such as the work of Dave et al. [2003] indicate better results for the combination of bigrams and trigrams.

2.1.3. Part of Speech. Some types of words are more likely to carry information about the polarity of a sentence or document, and, hence, part of speech can be a good discriminator in order to detect such words. It is indicated in previous works that adjectives are very important in determining the sense of the text. In fact, adjectives can be used both as main features, such as in works by Mullen and Collier [2004] and Whitelaw et al. [2005], and as filters for selecting other features. For instance, Turney uses adjectives to detect a set of phrases as features and then determines the polarity of documents based on those features [Turney 2002].

In addition to adjectives, other part-of-speech tags such as nouns like "gem" or verbs such as "love" can improve the performance of the task [Pang and Lee 2008]. Some previous works focus on comparing the effectiveness of adjectives, adverbs, verbs, and nouns in the classification task, including Benamara et al. [2007], Nasukawa and Yi [2003], and Wiebe et al. [2004].

2.1.4. Syntax. Several researchers investigate usage of dependency-based features by using dependency trees [Liu 2011]. There are contradicting results regarding the effectiveness of dependencies in text in previous works. Slight improvements in performance are reported in Dave et al. [2003], Gamon [2004], while Ng et al. [2006] conclude that addition of dependency-based features does not offer any improvements over the simple n-gram-based classifier.

2.1.5. Negation. The use of negating words in a sentence may totally flip the polarity of that sentence. For instance, ignoring "not" in "He does not like the color blue" results in a false positive. Attaching "not" to the words occurring near the negating words is one of the elementary techniques done for the first time by Das and Chen [2001]. Although the naïve assumption that each negation word flips the polarity of a window of following words is working in many cases, it is not a general rule. Later works try to optimize this technique by reversing the polarity of the phrases based on the part-of-speech tag patterns [Na et al. 2004].

Besides the explicit negation words, there are other terms that may negate a sentence. For instance, the verb "prevent" in the sentence "They prevent keeping unhealthy foods in the store" and the verb "deny" in "She denies admiring the brand" are implicitly reversing the polarity.

2.1.6. Topic-Oriented Features. Sentiment of a given sentence may be topic specific. For instance, the word "fast" in the context of car reviews is considered as *positive*, while it may be considered as *negative* in movie reviews. Different features are investigated based on topic in the literature specially in the work of Mullen and Collier [2004].

2.2. Works on Unannotated Data

It is obvious that coming up with a solution for unannotated data is always harder because of the lack of labels compared to annotated ones. In fact, most of the informative and also subjective text formats, such as comments, reviews and news, are left unlabeled, and hence there is no avail to using them for the purpose of training a classifier.

Researchers try to tackle the problem of unlabeled data from a wide range of perspectives. We have categorized the related methods in three different groups. The first group of solutions aims to expand a lexicon of words that contains words and their prior polarity and are explained in Section 2.2.1. The second group is concerned with domain adaptation, which is described in Section 2.2.2. Most of the works on unannotated data

would fall into one of these two categories. However, there are also many other types of works done in this scope. We will describe these methods in Section 2.2.3.

2.2.1. Lexicon Expansion. A very basic and simple idea to build a classifier for unannotated data is to use a lexicon of words. A lexicon is a dictionary of words, each word associated with a score showing its degree of polarity. If it is developed for emotion-mining purposes, then it may show the degree for each of the possible emotions. On classification time, polarity scores of each word contained in a test sample are fetched and processed in order to predict the polarity of the whole text. The processing of these scores could be done in different ways, including summing up, taking the average, and so on. This generic solution is called a "lexical-based" method. Currently, existing lexicons can be used for this purpose; however, to have higher performance, one may need to create his/her own lexicon of words suitable for the domain in question. Since manually building a lexicon is a tedious and time-consuming task, automatic solutions, called "lexicon expansion" methods, are suggested. Researchers apply different methods for automatic creation of the lexicon from the information lying in the data. This type of method, which expands the lexicon based on the information in the corpus, is called "corpus-based lexicon expansion." The first works belong to Hatzivassiloglou and McKeown [1997], Hatzivassiloglou and Wiebe [2000], Pang et al. [2002], and Yu and Hatzivassiloglou [2003]. They approached the problem by making simple assumptions about the occurrences of words. For instance, Pang et al. [2002] assumed that words present near the word "excellent" could be counted as *positive* while words adjacent to the word "poor" can be *negative*. In general, the potential words, whom the lexicon expansion is initiated with, are called "seed words."

Further attempts to create a useful lexicon were concerned with clustering of words or phrases in sentiment clusters including the works of Andreevskaia and Bergler [2006], Esuli and Sebastiani [2005, 2006a, 2006b], Finn and Kushmerick [2006], Takamura et al. [2007], and Kaji and Kitsuregawa [2007]. One of the good attempts in this set of works was the work of Hatzivassiloglou and Wiebe [2000]. They created a lexicon by using "opposition constraints" such as "but" and "and" between pairs of words and thereafter clustered the words to two partitions.

After finding the clusters of the words and in order to assign sentiment orientation (or degree of polarity) to them, different techniques have been proposed. For instance, Hatzivassiloglou and Wiebe [2000] simply assume that the words with more frequency seem to be *positive*.

Another popular technique is to have a set of seed words with their polarity and then to assign the polarity of new words with respect to their relationship to the seed words. In other words, polarity of the new words are assigned by propagating the polarity of seed words (based on the clustering results) such as the work of Andreevskaia and Bergler [2006], Gamon and Aue [2005], Esuli and Sebastiani [2005, 2006a], and Kamps et al. [2004]. This category of methods is called "dictionary-based lexicon expansion."

It is worth pointing out that most of the mentioned methods try to find a "prior polarity" of words. The "prior polarity" of a word is the polarity that it invokes no matter what context that word is occurring in, while "contextual polarity" is the polarity of the word with respect to the context. For instance, since the word "security" bears a *positive* polarity in general, we can assume a *positive* prior polarity for it. However, if it occurs inside a sentence like "There are three living former Secretaries of Homeland Security," then it does not infer any *positive* or *negative* polarity, since it is part of a referring expression. Therefore it has a *neutral* "contextual polarity." Prior polarity should be further applied to determine the "contextual polarity" of the words with respect to the concept and domain such as in the work of Wilson et al. [2005b].

2.2.2. Domain Adaptation. One idea to produce a generic classification method that is adaptable to any kind of data on any domain and extremely useful for unannotated data is training a classifier over a labeled dataset from one domain or topic, called the "source," and use it to label the unlabeled data from another domain, called the "target." However, results of doing so in various domains is shown to be unsatisfactory [Blitzer et al. 2007]. This is expected, since the keywords and phrases used in one domain may differ totally from the keywords in another one. Furthermore, one word in a domain may bear a different sentiment from what it does in another domain. Therefore, adapting the classifier trained over the source to be useful for the target is an essential step. This procedure is called "domain adaptation."

According to Jiang and Zhai [2007], domain adaptation is considered in two distinct attitudes, namely "labeling adaptation" and "instance adaptation." In labeling adaptation, the labeling function is adapted, since some features (words in opinion mining) may differ in polarity between source and target domains. In instance adaptation, the probability function of features is adjusted; for instance, the changes of word frequency from one domain to another one are modeled.

Early attempts to approach the problem relates to the work of Aue and Gamon [2005], in which they evaluate the performance of four rudimentary approaches to somehow adapt a classifier to be useful for the target domain. Those approaches include the following: training over all possible domains, limiting features to those observed in the target domain, ensemble of classifiers, and using a small set of labeled in-domain data.

Further simplistic attempts are the work of Yang et al. [2006], in which they ranked features of the two labeled datasets by running logistic regression over the sentences and selecting the highly ranked features as the ones that are most common in all domains.

Label transferring is another methods used in some of the previous works for domain adaptation. The basic idea of label transferring is to find the most informative samples of the target domain by means of a classifier that is trained over the source domain and then label those informative instances to train a brand new classifier over them. The first work that exploited this idea is Tan et al. [2007]. Later, the same team improved the performance of their system with selecting "generalizable features" by means of a measure they named "Frequently Co-occurring Entropy." Recently, Li et al. [2013] applied the same idea by finding the most informative instances in the target domain using classifiers with a query by committee strategy.

A very common technique, used in different schemes in previous works, is to cluster the features in every domain into two groups. The first group belongs to features that, regardless of the domain, happen frequently, called "domain independent." The second group, called "domain-specific" features, are common just inside their belonging domain. The reason to do such a clustering is to somehow align domain-specific features of the source domain to those of the target domain and then adapt the trained classifier in the source domain.

Based on the explanation above two steps should be followed:

Clustering features: There are methods suggested to distinguish the two types of features. The idea to recognize domain-independent features is to find features that occur more than a threshold in any domain. To find domain-specific ones, the degree of dependency of each feature to each domain should be calculated. In information theory, this can be done by using "mutual information" between the feature and domain.

Alignment: Alignment is a step in which each domain-specific feature in the target domain is mapped to one or more domain-specific features in the source domain. In the literature, this is done in various ways. In the first attempts, Blitzer et al. [2007] approached the problem by using an algorithm, called "structural correspondence learning" (SCL). SCL tries to find the domain-independent features (pivot features) as the

most frequent ones and finds the correspondence model between pivot features and all other features by training linear pivot trainers.

Li et al. [2009] try to approach the alignment problem by using non-negative matrix tri-factorization of term-document matrix. Basically, they factorize the term-document matrix in the source domain and then, by means of a matrix (that expresses if each of the words is occurring in both domains or not), estimate the factors of the term-document matrix in the target domain.

Pan et al. [2010] aimed to cluster domain-specific words in both domains by means of a spectral feature alignment algorithm. This work is promising to exploit all the relationships between the domain-specific and domain-independent words in spite of SCL. Basically, they create a bipartite graph of features that consists of two clusters of domain-specific and domain-independent features. Then, if there exists two domain specific features from two domains that have a lot of common domain-independent features, they align them to be in correspondence to each other.

In addition to the clustering-alignment method, there exists another group of solutions to the problem of adaptation that is based on feature selection in both of the source and target domains. This approach tries to find a feature space in which the gap between the distribution of source and target domains is minimum, comparing to other spaces. Features of both domains are transferred to this new feature space, and then a classifier over the source domain in the new feature set is trained. This classifier can be guaranteed to be working with higher performance over the target domain.

2.2.3. Other Methods. Some other methods appropriate for unannotated data include, but are not limited to, the following.

Bootstrapping: The general idea is to use an initial pre-trained classifier on another dataset to label the target dataset and then use this newly labeled dataset to train a new classifier. Kaji and Kitsuregawa [2006] use this method to label a set of HyperText Markup Language (HTML) documents with the *positive/negative* polarities.

Belief network modelling: One of the recent usages of belief networks is on training a model for the task of sentiment classification. Lin and He [2009] add a layer of sentiment to the structure of a famous probabilistic document model called "Latent Dirichlet Allocation" (LDA) [Blei et al. 2003] to find the polarity of words inside a set of documents with respect to each topic.

Combining lexical and machine-learning methods: Lexical and learning methods can be combined to compensate the disadvantages and drawbacks of each other. In order to optimize the performance of an initially trained classifier (over a different domain), Qiu et al. [2009] use a lexicon-based classifier in which in each step first the lexical classifier labels the data and then the learning classifier is trained over the labeled dataset. Operations continue until results of the two datasets have the minimum distance. In another work, Prabowo and Thelwall [2009] try to build a semi-supervised hybrid classifier by using both rule-based classifiers and SVM classification.

There are other works in which the task of classification was not completely based on the raw words. For instance, Hu et al. [2013] use emoticons to find the sentiment of a given comment in social media.

3. POLARITY RELATED RESOURCES

Research and analysis of the polarity classification methodologies requires resources such as lexicons and annotated datasets. One might need to generate his/her own resources by manually labelling them. Annotating sentiment of the textual data can be a tedious and time-consuming task for an individual. Also, since sentiment of a text is a subjective matter and is interpreted differently among various audiences, it is necessary to have more than one annotator to incorporate multiple perspectives in the

Table I. Summary of Polarity-Related Lexicons

Name	Author	Year	Size (words)	Set of polarities
Harvard General Inquirer	P. Stone	1968	11,790	positive, neutral and negative
Opinion Lexicon	B. Liu	2005	6,786	positive, negative
MPQA	T. Wilson	2005	8,222	positive, neutral, negative and both
WPARD	D. A. Medler	2005	1,400	positive, negative
SentiWordNet 3.0	S. Baccianella	2010	155,287	degree of polarity
NRC	S. M. Mohammad	2009+	various sizes	positive, negative

annotation. Although crowdsourcing tools are an option for data annotation, utilizing them might lead to a poorly annotated resource, since the annotators contributing in these tools are mostly regular people with no knowledge in areas such as psychology, linguistics, and sociology.

The challenging nature of the sentiment annotation encourages most of the researchers to take advantage of currently existing resources. Even if there exists an annotated dataset for the domain of the research, it is still time consuming to find it. Here we introduce some of the most well-known lexicons and datasets for polarity mining. Useful resources for emotion mining are explained in Section 6. Note that getting to know the process for the creation of these resources helps if one desires to build his/her own lexicon or dataset.

3.1. Lexicons

There are many publicly available lexicons that are results of lexicon creation and expansion of the previous sentiment analysis works. Among these lexicons, the following ones are known to be the most frequently used and effective in the literature (a summary for the following lexicons can be seen in Table I).

3.1.1. Harvard General Inquirer. This lexicon[1] is the result of one of the first attempts [Stone et al. 1968] to compile a list of words for sentiment analysis. The lexicon contains syntactic, semantic, and pragmatic information of its words. Among the information provided for each word, the one that is of interest is "*positive*" and "*negative*." The lexicon includes 11,790 words. The score of each word in this lexicon would be 1, 0, or −1, meaning that the word is *positive, neutral*, or *negative*, respectively.

3.1.2. Opinion Lexicon. This lexicon,[2] which is an outcome of Bing Liu's research in sentiment analysis [Hu and Liu 2004; Liu et al. 2005], consists of 6,786 words among which 2,009 of them are *positive* and the rest are *negative*. The corpus from which they have extracted the words includes customers' opinions about various features of products. They have extracted the words by finding sentences that include a frequent feature of a product and pulling adjectives from those sentences. Afterwards, they have separated those extracted words into two clusters of *positive* and *negative* ones based on the score of their synonyms and antonyms using a dictionary of words. The score of each word is defined in a similar way to the scoring of the Harvard General Inquirer lexicon.

3.1.3. Multi-perspective Question Answering (MPQA). The "MPQA" lexicon[3] [Wilson et al. 2005b] consists of 8,222 words, each of which are provided with a set of information including how subjective the word is; how strong its subjectivity is; the prior polarity

---

[1]http://www.wjh.harvard.edu/∼inquirer/spreadsheet_guide.htm.

[2]http://www.cs.uic.edu/∼liub/FBS/sentiment-analysis.html#lexicon.

[3]http://mpqa.cs.pitt.edu/lexicons/subj_lexicon/.

of the word, which can be *positive, negative*, *both*, or *neutral*; and whether the word is stemmed.

This lexicon is built on top of a subjectivity lexicon that resulted from the works of the same team [Wilson et al. 2005a]. In the first step, annotators are given a set of instructions and an annotating scheme to annotate phrases and words to be *positive, negative, both*, or *neutral*. In the second step, they measure the agreement between annotations of two annotators to evaluate the lexicon. Based on their annotation scheme, annotators' decisions depend mostly on the emotion of the sentences inside their corpus. This can make the MPQA lexicon a beneficial lexicon both for emotion classification and polarity classification.

3.1.4. WPARD. Using an online form, Medler et al. [2005] collected information from 342 undergraduate students. Participants were asked to rate how *negative* or *positive* were the emotions they associate with each word, using a scale from −6 (*very negative*) to +6 (*very positive*). They built the lexicon Wisconsin Perceptual Attribute Rating Database (WPARD)[4] from these data such that each word has a corresponding polarity and a real number showing the strength of that polarity.

3.1.5. SentiWordNet 3.0. "SentiWordNet 3.0"[5] is a lexical resource provided by Baccianella et al. [2010]. This lexicon is built on top of its previous version, SentiWordNet 1.0. It contains 155,287 words and is provides each word with a decimal signed polarity degree. In a nutshell, their method to create SentiWordNet 3.0 consists of five steps, including starting from a seed set of *positive* and *negative* words and applying synonyms and antonyms to expand the lexicon, adding objective words as a new cluster, training a community of ternary classifiers for the glosses of the words, classifying clusters of words with the classifiers, and, finally, running a random walk on the graph of words to make their scores converge to a final state. Recently, efforts have been made to adapt SentiWordNet to other languages. For example, Das and Bandyopadhyay [2010] develop SentiWordNet for three Indian languages (Bengali, Hindi, and Telugu) and Vu and Park [2014] construct a Vietnamese version of SentiWordNet.

3.1.6. NRC. Starting from 2009 to now, S. M. Mohammad has compiled several word-sentiment lexicons[6] from different corpora, including Twitter and customer reviews of Yelp and Amazon. In some cases, labeling is done manually and in other cases it is done automatically, such as using hashtag of a tweet as its label. For more elaboration on each of them, one can refer to Svetlana Kiritchenko and Mohammad [2014], Mohammad et al. [2013], Zhu et al. [2014], and Kiritchenko et al. [2014].

3.2. Datasets

Compared to other areas of text categorization, including emotion classification, polarity classification benefits from a larger number of well-annotated datasets. Because of this, here we only point to the benchmark datasets that have been very commonly exploited in the literature for experiments.

3.2.1. Amazon. "Amazon"[7] is the result of the work of Blitzer et al. [2007]. It is a dataset of product reviews constructed from the Amazon website. It includes four different domains of DVDs, books, electronics, and kitchen items, each of which has 2,000 reviews. The reviews of each domain are half *positive* and half *negative*, making this dataset balanced. Each instance in this dataset includes detailed information of a

---

[4]http://www.neuro.mcw.edu/ratings/.
[5]http://sentiwordnet.isti.cnr.it/download.php.
[6]http://saifmohammad.com/WebPages/lexicons.html.
[7]http://www.cs.jhu.edu/∼mdredze/datasets/sentiment/.

review consisting of its rating, which is from 0 to 5 stars, review title and date, and the review content. They have crawled the data from the Amazon website, annotated the reviews such that ratings higher than 3 stars are *positive* and those with ratings lower than 3 stars are *negative*, and discarded the ones with 3 stars, since those reviews are more likely to have ambiguous sentiments. In addition to the labeled data, this dataset includes 3,685 unlabeled instances in the DVD domain and 5,945 unlabeled instances of kitchen reviews. This part of the dataset is created by selecting an equal number of *positive* and *negative* reviews from a set of labeled data and discarding the labels.

3.2.2. Movie Datasets. Various versions of datasets[8] are extracted from the movie reviews of famous online movie databases, all of which are built by Pang et al. Here is a summary of each version:

—Pool of HTML files: These data consist of 27,886 HTML files that are unprocessed and unlabeled. Files consist of reviews crawled from an online database called "Internet Movie Database" (IMDB). This is the raw version of the next labeled one (Polarity dataset).
—Polarity dataset: This version of the data includes four different subversions. Subversions 0.9 and 1 [Pang et al. 2002] consist of 700 *positive* and 700 *negative* processed reviews, and subversion 1.1 is slightly modified by removing a few non-English/incomplete reviews and correcting some mislabeled reviews. Finally, the last subversion 2 consists of 1,000 reviews for each class of polarities [Pang and Lee 2004]. Since not all the reviews in the raw version have the same format of rating, labeling them is done differently based on the format of rating. First, only those reviews whose author has explicitly declared the rating are classified. With a 5-star system, reviews with 3.5 stars and up are labeled as *positive* and reviews below or equal to 2.5 are counted as *negatives*. With a four-star system, reviews higher or equal to 3 stars are labeled as *positive*, and the ones with 1.5 stars or lower are labeled as *negative*. Finally, with a letter grade system, B or above is considered *positive* and C or below is considered *negative*.
—Sentence polarity dataset: This version of the data includes 5,331 *positive* and 5,331 *negative* processed sentences and snippets provided by Pang and Lee [2005]. All of the instances are downloaded from a movie review database called "rottentomatoes," which classifies reviews either as fresh (meaning *positive*) or as rotten (meaning *negative*).

3.2.3. Blogs. This dataset, which is provided by Melville et al. [2009], includes two different sets of blog posts, one of which is concerned with technology blogs, and the other one is related to political blogs. The first set, named "lotus blogs," is a set of posts corresponding to IBM Lotus collaborative software gathered from 14 blogs, 4 of which have posted mostly *negative* comments about the product, and the others have provided *positive* posts. The data were provided by downloading either the latest posts of each blogger's Rich Site Summary (RSS) feed or the archived posts of that blog. Afterwards they extracted text from those parts of the HTML files in which the ratio of tags to words is above a minimal threshold. Then all the posts were read and labeled manually to be *positive*, *negative*, *neutral*, or *irrelevant*. There exist 34 *positive* and 111 *negative* instances in this set.

The second part of this dataset consists of political posts regarding two candidates of the United States presidential election in 2008, namely "Barak Obama" and "Hillary Clinton." The posts were taken from a set of 16,741 blogs, filtered based on whether they have the words "Obama" and "Clinton," and randomly selected for manual labeling.

---

[8]http://www.cs.cornell.edu/people/pabo/movie-review-data/.

Table II. Summary of Polarity-Related Datasets

Name	Author	Year	Size	Type of Data
Movie	B. Pang	2004	29,419	processed labeled IMDB movie reviews (document level)
		2004	2,000	raw unlabeled IMDB movie reviews (document level)
		2005	10,662	processed rottentomatoes movie reviews (sentence level)
Amazon	J. Blitzer	2007	8000	reviews of products
Blogs	P. Melville	2009	252	product review and political posts

Based on Melville et al. [2009], labeling political posts is much more difficult than that of product reviews, because posts are more emotional, mostly mentioning implicit comments and judgments about the candidate, and may apply cultural references to make a point. Therefore they have labeled those posts that have explicitly mentioned an opinion about one of the two candidates as *positive* or *negative*. Hence there are no *neutral* or *irrelevant* posts in this set. The Politic dataset includes 49 *positive* and 58 *negative* posts.

4. SURVEY ON EMOTION THEORIES

In Sections 2 and 3, we discussed methods and resources for polarity classification that can almost equally be effective for emotion classification. In any emotion-related research, the first question to be answered is what emotion really is. In this section, we introduce some theories that define emotion and suggest some sets of basic emotions. While most of the research on emotions in computer science uses the terms *emotion*, *feeling*, *mood*, and *affect* interchangeably, these terms do not share the same exact meaning. According to Fox [2008], in affective neuroscience, the terms are defined as follows:

—**Emotion:** discrete and consistent responses to internal or external events that have a particular significance for the organism; emotion has short-term duration.
—**Feeling:** a subjective representation of emotions, private to the individual experiencing them; similarly to emotion, it has short-term duration
—**Mood:** a diffuse affective state that compared to emotion is usually less intense but with longer duration
—**Affect:** an encompassing term used to describe the topics of emotion, feelings, and moods together.

Even with having clear definitions of these terms, there are still some controversial issues regarding whether some particular human states are classified as an emotion. For instance, *thankfulness* or *gratitude* is considered an emotion by some theorists, while others consider actions such as greeting, thanking, and congratulating as *communicative functions*.

Scientific studies on the classification of human emotions date back to the 1960s. There are two prevalent theories in this field: *discrete emotion theory* and *dimensional model*. *Discrete emotion theory* states that different emotions arise from separate neural systems, while the *dimensional model* states that a common and interconnected neurophysiological system is responsible for all affective states. This model defines emotions according to one or more dimensions where usually one of them relates to intensity of emotions.

Basic emotions refer to those that do not have any other emotion as constituent parts. In addition, they can be recognized by humans all over the world regardless of their race, culture, and language. Theorists of both sides have proposed sets of emotions that tend to be basic ones. Table III shows some of the frequently referenced models of basic emotions. Ekman, one of the earliest emotion theorists, suggested that those certain emotions that are universally recognized form the set of basic emotions. He later

Table III. Different Models of Basic Emotions Proposed by Theorists

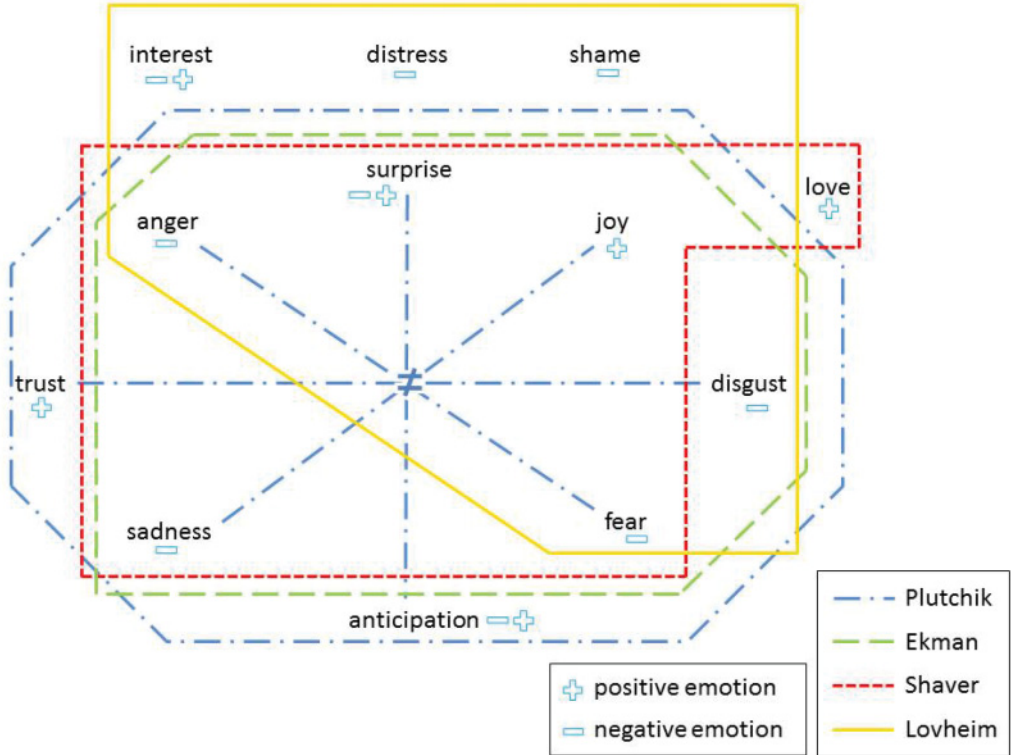| Theorist	Year	Basic Emotions	Type
Ekman	1972	anger, disgust, fear, joy, sadness, surprise	discrete
Plutchik	1986	anger, anticipation, disgust, fear, joy, sadness, surprise, trust	dimensional
Shaver	1987	anger, fear, joy, love, sadness, surprise	discrete
Lovheim	2011	anger, disgust, distress, fear, joy, interest, shame, surprise	dimensional

Fig. 2. The illustration of four frequently used emotion models.

expanded his set of emotions by adding 12 new *positive* and *negative* emotions [Ekman 1992]. The dimensional model of Plutchik and Kellerman [1986] arranges emotions on four bipolar axes: *joy* vs. *sadness*, *anger* vs. *fear*, *trust* vs. *disgust*, and *surprise* vs. *anticipation*. The fact that some of these emotions are actually opposite of each other is trivial in cases like *joy* vs. *sadness*, but it is not intuitive enough in other cases, such as *anger* vs. *fear*. Shaver et al. [1987] model emotions in a tree structure such that basic emotions are the main branches and each branch has its own categorization. Lövheim [2012] also suggests a dimensional model; however, his model differs from Plutchik's. Lövheim believes that three hormones, serotonin, dopamine, and noradrenaline, form three dimensions of a cube, where each basic emotion is placed on one of the corners.

Figure 2 illustrates the four explained models together so one can compare them. The Plutchik's bipolar division of emotions is shown using the sign $\neq$. The *positiveness* and/or *negativeness* of emotions are also shown using the signs + and −, respectively. Emotions such as *interest*, *surprise*, and *anticipation* can be both *positive* and *negative*, depending on the situation in which they are felt. Alm and Sproat [2005] even divide *surprise* to two separate emotions of *positively surprise* and *negatively surprise*. Table IV

Table IV. Commonality of Emotion Models

Emotion	Ekman	Plutchik	Shaver and Parrott	Lovheim
Anger	✓	✓	✓	✓
Anticipation		✓		
Disgust	✓	✓		✓
Distress				✓
Fear	✓	✓	✓	✓
Interest				✓
Joy	✓	✓	✓	✓
Love			✓	
Sadness	✓	✓	✓	
Shame				✓
Surprise	✓	✓	✓	✓
Trust		✓		

shows another illustrations of commonality of these emotion models. According to both Figure 2 and Table IV, *anger*, *fear*, *joy*, and *surprise* are common in all models, but there is no agreement on the rest. One interesting point in all models is that the number of *negative* emotions outweighs the number of *positive* ones. While psychologists do not agree on what model describes more accurately the set of basic emotions, the model suggested by Ekman et al. [1972], with six emotions, is the most widely used in computer science research.

5. EMOTION-MINING METHODOLOGIES

In this section, we explain the major works on textual emotion mining in the computer science world; however, note that research on emotion has been an interesting topic in many other fields as well. Murphy et al. [2015] studies the use of language to convey emotional experience, and Pennebaker [1997] investigates the effect of expressing emotions on physical and mental health. Recently, interdisciplinary studies among psychology, linguistics, computer science, and other areas has increased. For instance, Russell et al. [2013] is a joint work by a group of anthropologists, linguists, and psychologists to discuss how emotions are conceptualised by people.

In 1992, Walther [1992] introduced the *Social Information Processing (SIP)* theory, which states that in order to convey relational information in computer-mediated communications, people use verbal clues instead of nonverbal clues that would have been used in face-to-face environments. Walther et al. [2005] later validated their hypothesis by conducting an experimental study and showed that affinity is expressed equally effectively in both face-to-face and textual styles. In addition, verbal clues carried a larger portion of relational information in communications via a computer medium. This simple theory can be a proof for the validity of a textual emotion-mining research topic.

Since most of the body of research in emotion mining is dedicated to emotion classification, we put more emphasis on this division too; however, note that other directions of this field, introduced in the taxonomy, are also being investigated.

Automatic classification of emotions can be categorized from different aspects, similarly to the categorization of polarity tasks that we did in Section 2. For example, it can be done in document level vs. sentence level or can use annotated data vs. unannotated data. Although we dedicated separate sections to annotated and unannotated data for polarity tasks, since most of the body of emotion research focuses on sentence level and annotated data, we consolidate the work together.

In a text environment, emotion analysis can be either from the writer's or from the reader's perspective. The former refers to emotions that the author had when he/she was writing the message, while the latter refers to a user's affective response to being

exposed to feelings evoked by an emotional text. Readers can further be divided into two groups: an individual reader or a group or society of readers, sometimes referred to as *social emotion detection*. Both writer and reader can feel the same emotion in some cases; however, it is not a general rule. A reader's point of view has attracted less attention in the literature; nevertheless, it has many applications, including helping authors to predict how their work will influence the audience or helping readers to retrieve documents that have content relevant to their desired emotion [Rao et al. 2014]. Examples of social emotion detection can be found in Mishne and De Rijke [2006] and Lei et al. [2014].

In some configurations, each sample (a document or a sentence) is assumed to have one single emotion, while sometimes the text can be multi-emotional, which means it can contain several emotions at the same time. An example of this situation is the short document "I was happy that it was my birthday yesterday. I was anticipating my family to throw me a party. however, nobody remembering it made me sad" which shows *joy*, *anticipation*, and *sadness* simultaneously.

Techniques used for polarity classification of both annotated and unannotated data, discussed in Section 2, are all prevalent methods in emotion classification as well. Therefore, we do not replicate them here and instead give a thorough review of existing methods specific to emotion classification with enough elaboration.

Hancock et al. [2007] is one of the works to characterize how users express emotions in text-based systems. Their study on 40 undergraduate male and female students showed that both genders agree more with their conversation partner when they want to convey a *positive* attitude. They also use 5 times less negative affect terms and use more punctuation marks. On the other hand, those partners who receive the emotional texts judge mostly based on negations and exclamation points. These findings are in line with what "SIP theory" suggests. This study contributes to automatic extraction of emotions from text by providing an insight into the strategies that people employ to convey their emotions.

Kao et al. [2009] is one of the earliest surveys on textual emotion mining. It classifies works into lexical-based (or keyword-based), learning-based, and hybrid methods where hybrids combine detecting keywords, learning patterns, and using other supplementary information. They then suggest a system in which keywords are extracted using a semantic analyzer, and an ontology is designed with the emotion theory of appraisal. These two are combined in a case-based reasoning architecture.

Jain and Kulkarni [2014] give a short survey on emotion-mining research but their review lacks a rational categorization of works. They introduce some Information Retrieval (IR) models that can be used in text research and suggest a system, called "TexEmo," that essentially uses a bag of words with Term Frequency-Inverse Document Frequency (TF-IDF) weighting as features and trains an SVM classifier on them. They do not report any results for this system.

Kim et al. [2010] follow lexical-based approaches to evaluate the merit of the "discrete emotion theory" and the "dimensional model," discussed in Section 4. To build a classifier based on the theory of discrete emotions, they use the Wordnet Affect lexicon as well as three-dimensional reduction techniques, namely Latent Semantic Analysis (LSA), Probabilistic LSA, and Non-negative Matrix Factorization. To build a dimensional classifier, they use a normative database of English affective words, called "Affective Norm for English Words," in which each word is rated on the three dimensions of valence, arousal, and dominance. According to their results on Semantic Evaluation (SemEval) 2007, International Survey on Emotion Antecedents and Reactions (ISEAR), and fairy tales datasets, all of which will be introduced in Section 6.2, performance of methods varies on each emotion, and there is no method that performs better than others on all emotions that are under discussion.

Alm et al. [2005] try to identify emotional passages and determine their valence (*positive* vs. *negative*). They extract 30 features from their dataset of children's fairy tales, including direct speech (if the sentence is a whole quote), punctuation marks, complete uppercase words, sentence length, range of story progress, and POS. Then, a linear classifier, called "Sparse Network of Winnows," is applied on the data. Although their classification results are unsuccessful, their dataset is reputed and widely used in the field of emotion mining.

Neviarouskaya et al. [2007] construct a rule-based system for emotion recognition, named "Affect Analysis Model" (AAM). They create an affect database that contains emoticons, acronyms, abbreviations, affect words, interjections, and modifiers. Each entry is manually labeled with an emotion and an intensity showing the degree of its affective state. This database is then used in a five-stage system: symbolic cue analysis, syntactical structure analysis, word-level analysis, phrase-level analysis, and, finally, sentence-level analysis. Each stage consists of a set of rules that help identify the emotion relied in the text. An example rule is as follows: "In a compound sentence that independent clauses are connected with comma, 'and', or 'so', the output emotion is equal to the emotion of the clause with maximum intensity." In a later work, Neviarouskaya et al. [2009] added the ability to process sentences of different complexity. To do so, they decompose a sentence to pieces that correspond to lexical units and then apply some extra rules to infer the total emotion of the text based on the emotions of its parts. AAM is claimed to handle informal messages and is tested on a dataset of diarylike blog posts; however, it still has a long way to prove this for other data. In addition, it cannot distinguish among different meanings of words with respect to the context and does not take into account the expression modifiers such as "*to death*" in the example "*I love my ipad to death*."

Chaumartin [2007] proposes another rule-based system, called "University Paris 7 (UPAR7)," specifically for the SemEval 2007 dataset. They use the Stanford syntactic parser to build the dependency graph for each news headline. Then they enrich the Wordnet Affect and SentiWordnet lexicons to use them for rating each word separately and then try to rate the main subject of the whole headline sentence, considering contrasts, accentuations, negations, modals, and so on. UPAR7 ranked as one of the top systems that competed in the SemEval 2007 category of shared task of affective computing.

Strapparava and Mihalcea [2008] predict emotions of news headlines in an unsupervised manner from the SemEval 2007 dataset. In one experiment, they use the LSA technique as a semantic similarity mechanism. Each document can be represented in an LSA space by summing up the normalized LSA vectors of all the terms contained in it. In another experiment, they train a Naïve Bayes classifier on a collection of LiveJournal blogs as a training set and use this classifier to label their news data. Their results are acceptable compared to three other algorithms that participated in the SemEval 2007 workshop.

Danisman and Alpkocak [2008] use a Vector Space Model (VSM) classifier in which each document is represented as a vector and each axis corresponds to a unigram word. The value of a word in a vector (a document) is calculated using TF-IDF. VSM is relying on two simplifying assumptions that documents with the same emotion form a contiguous region and a region of one emotion does not overlap with the others'. Having this model, on classification time, the test document is converted to a vector and the cosine angel between this vector and all other vectors in the model determines the similarity. They show that VSM outperforms SVM and Naïve Bayes classifiers on the SemEval 2007 dataset.

Gupta et al. [2013] use an algorithm from the boosting family, namely "Boostexter," that was initially proposed in Schapire [1999]. Each base classifier in Boostexter assigns a confidence value in addition to its prediction for each instance. For a test instance, the final classifier outputs the sum of all confidences of all classifiers per class.

They also show the effectiveness of using a set of "salient features" that are essentially some linguistic clues from a dataset of customers' emails to the customer service department of some companies. These salient features include negative emotions, negative opinions, and other expressions specific to the domain of customer care such as threats to take their business elsewhere, and so on. According to their results, adding salient features to traditional n-gram features improves the performance significantly.

Following a psychologically based approach, Ho and Cao [2012] use a high-order Hidden Markov Model (HMM) to address the emotion classification problem on the ISEAR dataset. They believe that emotion is the result of a sequence of mental states. Their idea is to transform the input text into a sequence of events that cause mental states and then automatically generate an HMM to model the process where this sequence of events causes the emotion. They get modest results over the four emotions of *anger*, *fear*, *joy*, and *sadness*, where *anger* includes both *anger* and *disgust*.

As stated in Section 4, "mood" is a less-intense state compared to emotion but has long-term effects. Mood classification, thus, is very similar to emotion classification and is partially addressed in the literature such as in G. Mishne's work [Mishne 2005]. Mishne [2005] attempts to classify blog posts into 1 of 40 moods, including *excited*, *sleepy*, *confused*, *crazy*, and so on. The author focuses mostly on feature selection by investigating the effectiveness of length-related and semantic-oriented features, frequencies of Part Of Speech (POS) tags, Pointwise Mutual Information (PMI) for each word and mood, and emphasized words. They believe that, due to the subjective nature of mood categories and annotations in the corpus, good results are not achieved.

5.1. Multi-Label Emotion Classification Research

In machine learning, multi-label classification algorithms are traditionally categorized into two classes: algorithm adaptation methods and problem transformation methods. The idea of the first approach is to adapt the existing single-label classification algorithm to enable it to classify multi-labeled data. In the second approach, using some transformation techniques, the multi-labeled data are transformed into another problem space, in which they have a single label and then a single-label classifier is applied on them [Bhowmick 2009]. In what follows, some of the multi-label emotion classifiers are introduced.

Given k different single labels, Bhowmick [2009] uses an ensemble-based approach, called "random k-label sets classifier," which basically consists of an ensemble of "Label Powerset" (LP) classifiers. Each LP learns one single classifier with k' possible labels, where $k' \leq k$ and is trained using a different small random subset of all emotions. A test instance is classified by combining votes from individual LP classifiers such that it is labeled with an emotion if the average vote of all classifiers is greater than a user-specified threshold. This work is an example of algorithm adaptation methods. Additionally, they explore the effectiveness of different feature sets such as polarity of subject, object, and verbs in sentences and semantic frame features using the Berkeley FrameNet lexicon [Baker et al. 1998]. Results of their experiments on a dataset of Indian news headlines reveal that the combination of polarity and semantic features is the best choice for a multi-label environment.

Luyckx et al. [2012] is another work on multi-label classification of emotional texts. They focus on a dataset of notes written by people who have committed suicide that is provided for track 2 of the medical Natural Language Processing (NLP) shared task, 2011.[9] The task is to predict label(s) of a note among 15 possible emotions, such as *hopelessness*, *love*, *pride*, *thankfulness*, and so on. We think that it is dubious to consider some of these labels, such as *instructions, information*, and so on, as emotions. First, they split all multi-labeled notes to single-labeled fragments manually. Then an SVM

---

[9]https://www.i2b2.org/NLP/Coreference/Call.php.

Table V. Summary of Current Emotion-Mining Methods

Name	Dataset	Emotions	Multi-label	Method
C. Alm et al. [2005]	fairy tales	categorizing anger, disgust, fear, joy, sadness, positive surprise, and negative surprise into positive, negative, and neutral	No	Sparse Network of Winnows
G. Mishne [2005]	LiveJournal	40 moods	No	Support Vector Machine
A. Neviarouskaya et al. [2007]	160 sentences from online blog posts	anger, disgust, fear, guilt, interest, joy, sadness, shame, surprise	No	Rule Based
F. R. Chaumartin [2007]	SemEval 2007	anger, disgust, fear, joy, sadness, surprise	No	Rule Based
C. Strapparava and R. Mihalcea [2008]	SemEval 2007	anger, disgust, fear, joy, sadness, surprise	No	(1) unsupervised: knowledge based, (2) supervised: Naive Bayes
T. Danisman and A. Alpkocak [2008]	SemEval 2007	anger, disgust, fear, joy, sadness	No	Vector Space Model
A. Neviarouskaya et al. [2009]	diarylike blog posts	anger, disgust, fear, guilt, interest, joy, sadness, shame, surprise	No	Rule Based
P. K. Bhowmick [2009]	Indian news headlines	disgust, fear, happiness, sadness	Yes	ensemble of Label Powerset classifiers
S. Kim et al. [2010]	SemEval 2007, ISEAR, and fairy tales	anger, fear, joy, sadness	No	unsupervised: lexical based
D. T. Ho and T. H. Cao [2012]	ISEAR	anger (including disgust), fear, joy, and sadness	No	Hidden Markov Model
K. Luyckx et al. [2012]	600 suicide notes for track 2 of the 2011 medical NLP challenge	instructions, hopelessness, love, information, guilt, blame, thankfulness, anger, sorrow, hopefulness, fear, happiness, peasefulness, pride, abuse, forgiveness	Yes	Support Vector Machine
N. Gupta et al. [2013]	set of 1,077 customers' emails	factual, emotional	No	Boosting
M. C. Jain and V. Y. Kulkarni [2014]	—	anger, disgust, fear, joy, sadness, surprise	No	Support Vector Machine

with Radial Basis Function is trained on these single-labeled data. Finally, a threshold is set for SVM's probability estimated for each emotion; if the probability exceeds the threshold, then that emotion is assigned to the sentence. Their method has improved the recall compared to a baseline method with the cost of degrading the precision.

Table V shows a summary of the explained methods in this section, in chronological order. They are compared with respect to the dataset and set of emotions they use, as well as the main characteristics of their approach.

5.2. Emotion Mining Research on Twitter

With more than 300 million active users and 500 million tweets per day,[10] Twitter is a popular network for sharing personal feelings and moods with acquaintances and

---

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

friends. Hence, significant research is devoted to Twitter data with the purpose of analyzing the emotions expressed in tweets. Being short and informal, having misspellings, and using hashtags, special symbols such as emoticons and emojis, short forms of words, and abbreviations are properties that discriminate tweets from normal texts and add to the complexity of the task.

Bollen et al. [2011] analyze emotions of all tweets in a specific time frame. They use a psychometric test, named "Profile of Mood States" (POMS) consisting of 793 adjective terms, each related to a particular emotion. Then the probability of each tweet showing an emotion is calculated based on these features, and the results are aggregated over all the tweets of 1 day. Finally, the overall emotions of tweets are compared with global events of that period and some correlations are found. Although this method does not consider the reader's perspective, it may still be classified as a social emotion detection task, introduced earlier in this section.

Hashtags are space-free phrases following the "#" character such as #mickeymouse and #iamhappy. They can be used as indexes to search for related content or grouping messages. Hashtags are widely used in Twitter as they convey valuable information in a short piece of text. Wang et al. [2012] build a dataset from Twitter, containing 2,500,000 tweets and use hashtags as emotion labels.[11] In order to validate this type of labeling, they select 400 tweets randomly and label them manually. Comparing manual labels and hashtag labels show acceptable consistency. Then they explore the effectiveness of different features such as n-grams, different lexicons, POS, and adjectives in detecting emotions. Their best result is obtained when unigrams, bigrams, lexicons, and POS are used. Finally, they show that increasing the size of the training set has a direct effect on accuracy. While their dataset is a good source of emotional tweets, it is highly imbalanced, and the use of some unclear hashtags as emotion labels, such as #embarrass for *sadness*, makes soundness of the dataset open to criticism.

Hasan et al. [2014] also validate the use of hashtags as emotion labels on a set of 134,000 tweets. To this end, they compare hashtag labels with labels assigned by a group of people as well as those assigned by a group of psychologists. They found that crowd labels are not in agreement even with themselves; however, psychologists' labels are more consistent and show more agreement with hashtags, too. Therefore, they cast doubt on the use of crowd labeling such as in Amazon's Mechanical Turk for tasks related to emotion mining. They also introduce a supervised classifier, named "EmoTex." It essentially uses the feature set of unigrams, list of negation words, emoticons, and punctuations and runs K-Nearest Neighbors (KNN) and SVM on the training data.

Roberts et al. [2012] create a corpus of 7,000 manually labeled tweets that are retrieved by searching for 14 emotion evoking topics, such as World Cup and Christmas. There are a total of seven emotions where each tweet can have zero, one, or many of them. Seven binary SVMs, one for each emotion and each with a different feature set, are trained. Features include n-grams, punctuation, hypernyms, and topics. To obtain topics, they assume that each tweet associates with a probabilistic mixture of topics, and they are inferred using LDA. Their best performance was over the emotion *fear*, which led them to infer that *fear* is highly lexicalized with less variation than other emotions.

Mohammad [2012] introduced his corpus, called the "Twitter Emotion Corpus" (TEC), collected from Twitter, that will be explained in Section 6.2 and, similarly to Roberts et al. [2012], built binary SVMs, one for each emotion, using unigrams and bigrams as features. He then showed the effectiveness of this corpus in cross-domain classifications by using these data to predict emotions on another dataset, SemEval 2007. He also built a lexicon from this corpus that will be introduced in Section 6.1.

---

[11]Their dataset is available for download at http://knoesis.org/projects/emotion.

Table VI. Summary of Current Emotion-Mining Methods on Twitter

Name	Dataset	Emotions	Method	Labeling Process
J. Bollen et al. [2011]	crawled about 9,000,000 tweets	tension, depression, anger, vigour, fatigue, Confusion	Profile of Mood States	no labeling
W. Wang et al. [2012]	crawled about 2,500,000 tweets	anger, fear, joy, love, sadness, surprise, thankfulness	linear classifier	using hashtags
K. Roberts et al. [2012]	crawled 7,000 tweets from 14 emotion evoking topics	anger, disgust, fear, joy, love, sadness, surprise	Support Vector Machine	manual
S. M. Mohammad [2012]	built TEC by crawling about 21,000 tweets	anger, disgust, fear, joy, sadness, surprise	Support Vector Machine	using hashtags
M. Hasan [2014]	crawled about 134,000 tweets	twoo-dimensional model: active, inactive/ happy, unhappy	Support Vector Machine and K-Nearest Neighbors	using hashtags
W. Li and H. Xu [2014]	16,485 posts from Weibo, a Chinese microblogging website	anger, disgust, fear, joy, sadness, surprise	Support Vector Regression	manual

Table VI depicts the summary of the explained methods working on Twitter data, sorted in chronological order. They are compared for the dataset and set of emotions they use, as well as the main characteristics of their approach.

5.3. Emotion Mining for Other Languages

Most of the work in textual emotion mining is on the English language; nevertheless, it is worth mentioning the few works done on other languages, since the ideas and techniques may still be used in a language-agnostic way.

Li and Xu [2014] try to detect emotions from messages in Weibo, a Chinese microblog website with functionalities thoroughly similar to Twitter. They believe that the accuracy of detecting emotions in a text can be increased if we look for the events that cause emotions. In this manner, their work is similar to Ho and Cao [2012]. Therefore, they adopt the notion of *cause events* that are meant to be the reasons of certain emotions. To spot cause events and use them as features, they exploit a marker list, containing keywords to mark the occurrence of cause events; an emotion list, containing keywords expressing emotions; and a linguistic pattern set, describing how emotions and cause events are arranged in a text. All of these resources are adapted to the informal environment of Weibo. Then a "Support Vector Regression" (SVR), an algorithm from the family of SVMs, is trained using these features. According to the results, performance is boosted for some emotions, although it is decreased for others, such as *fear* and *sadness*. Lei et al. [2014] is another example of an emotion-mining study in Chinese that will be explained in Section 6.1. Also, the aforementioned method of Bhowmick [2009] has addressed the emotion-mining task on an Indian dataset in a multi-label environment.

In addition, since most of the tools for emotion mining are built for the English language, a portion of the works are dedicated to providing resources specific for other languages by either developing a resource from scratch or adapting existing English resources. Examples of adapting SentiWordNet for the Indian and Vietnamese languages are presented in Section 3.1.5.

Table VII. Summary of Emotion-Related Lexicons

Name	Author	Year	Size (words)	Set of Emotions
Wordnet Affect	C. Strapparava	2004	4,787	a hierarchy of emotions
LIWC	J. W. Pennebaker	2007	5,000	affective or not, positive, negative, anxiety, anger, sadness
NRC	S. M. Mohammad	2010	14,182	anger, fear, anticipation, trust, surprise, sadness, joy, disgust
NRC hashtag	S. M. Mohammad	2013	32,400	anger, fear, anticipation, trust, surprise, sadness, joy, disgust
CBET	A. Gholipour Shahraki	2015	24,000	anger, fear, joy, love, sadness, surprise, thankfulness, disgust, guilt

6. REVIEW OF EMOTION-RELATED RESOURCES

6.1. Lexicons

Almost all of the emotion-mining works rely on using a lexicon. Lexicons are very useful in that they give prior information about the type and strength of emotion carried by each word or phrase. In this section, we introduce some of the lexicons useful for the emotion-mining task. Their characteristics are summarized in Table VII.

6.1.1. Wordnet Affect. Wordnet Affect[12] is an emotional lexical resource, including a list of sets of synonym words, referred to as *synsets*. The set of emotions in this lexicon is hierarchically organized. Strapparava and Valitutti [2004] build this lexicon on top of their previous lexicon, Wordnet. They manually form an initial set of 1,903 affective words and expand them by adding their corresponding nouns, verbs, adjectives, adverbs, and so on. Then a subset of synsets of Wordnet that contain at least one of these affective words are selected, and the rest are rejected. This forms the core of the lexicon. Then the lexical and semantic relations between synsets of this core lexicon and other synsets of Wordnet are examined to see if they preserve the affective meaning represented by those core synsets. After adding new synsets, Wordnet Affect contains 2,874 synsets and 4,787 words. One interesting feature of this lexicon is the notion of stative/causative for words. A word is causative if it refers to an emotion that is caused by that entity (e.g., amusing). On the other hand, a word is said to be stative if it refers to the emotion owned or felt by that subject (e.g., amused).

6.1.2. LIWC. The Linguistic Inquiry and Word Count (LIWC)[13] is another emotion-related lexicon developed by Pennebaker et al. [2007]. In the first step of generating this lexicon, some initial category scales are generated in a psychological process and then various scales are added to initial lists by brain-storming sessions. In the next step, three independent judges rate the words in two phases, such that after completion of each phase, all category scale lists are updated according to judges' rates. The initial LIWC judging took place in 1992 and, since then, it has been updated and largely expanded.

6.1.3. NRC. Mohammad and Turney [2010] develop the NRC word-emotion association lexicon.[14] Using Amazon's Mechanical Turk, they asked Turkers to annotate words, from non-specific domains, according to the emotion they evoke. One important challenge in this process is malicious annotations that can happen in cases where words in different senses evoke different emotions. To solve this problem, the target

---

[12]http://wndomains.fbk.eu/wnaffect.html.
[13]http://www.liwc.net/.
[14]http://www.saifmohammad.com/WebPages/lexicons.html.

sense needs to be conveyed to annotators. Hence, they asked additional questions from Turkers, including word choice questions, that help identify instances where the annotator may not be familiar with the target term. In addition to building a lexicon, they concluded that a regular crowd can produce reliable emotion annotation, given proper guidelines. This is in contrast with findings of Hasan et al. [2014], who showed that crowd labeling of emotional tweets have quite low inter-agreement with each other and with emotional hashtags of tweets.

6.1.4. NRC Hashtag. In another attempt, the main author of the NRC lexicon, S. M. Mohammad, developed another useful lexicon, called the "NRC hashtag emotion lexicon"[15] [Mohammad 2012]. Using a corpus of 21,000 tweets (TEC), the Strength of Association (SoA) for an n-gram n and an emotion e is calculated to be

$$SoA(n, e) = PMI(n, e) - PMI(n, \neg e), \tag{1}$$

where *PMI* is the pointwise mutual information, calculated as

$$PMI(n, e) = log \frac{freq(n, e)}{freq(n) * freq(e)}, \tag{2}$$

where $freq(n, e)$ is the number of times that n occurs in a tweet that has the label e, and $freq(n)$ and $freq(e)$ are hte frequencies of n and e, respectively, in the corpus. $PMI(n, \neg e)$ is calculated likewise. Words having SoA greater than zero are kept in the lexicon.

6.1.5. Clean Balanced Emotional Tweets (CBET). This lexicon[16] is compiled by Gholipour Shahraki [2015] from the single-labeled part of a dataset with the same name. This dataset contains a large number of tweets, each labeled with one single emotion (for more information about it, see Section 6.2). The lexicon is actually a $V \times E$ matrix, where V is the set of all the single words (unigrams) contained in CBET dataset and E is the set of emotions covered in it. The element at index (j, i) of the matrix denotes the degree that word w_j expresses emotion e_i. In other words, each entry of the lexicon has a corresponding weight vector that contains weights associated to each of the participating emotions. The weight $F(e_i|w_j)$ is calculated as the number of times that w_j has occurred in tweets that have the label e_i in the dataset. That is,

$$F(e_i|w_j) = \sum_{s \in S} F(e_i|s) \times I_s(w_j), \tag{3}$$

where $F(e_i|s)$ is the presence of emotion e_i given tweet s and $I_s(x)$ is an *indicator function* that is equal to 1 if $x \in s$ and is 0 otherwise. The naïve assumption supporting this idea is that all the words in a tweet are in agreement with the label of that tweet. The CBET lexicon is the newest emotion lexicon; it is publicly available and covers more emotions compared to all previous ones.

In addition to these publicly available lexicons, there are other lexicons generated for specific tasks that are not accessible; nevertheless, reviewing their method of generation can still provide some ideas if one wants to build his/her own special-purpose lexicon.

6.1.6. Word-Emotion Mapping Lexicon. Katz et al. [2007] create a word-emotion mapping from the SemEval 2007 dataset that will be introduced in Section 6.2. A weight vector is assigned to each lemmatized word w from the corpus such that each element in this vector corresponds to one emotion. The value of this element then is calculated to be the average emotion score observed in all samples in which w participated.

---

[15]http://www.saifmohammad.com/WebPages/lexicons.html.
[16]http://www.cs.ualberta.ca/~zaiane/data/CBET.

Table VIII. Summary of Emotion-Related Datasets

Name	Author	Year	Size	Type of Data
ISEAR	K. R. Scherer	1997	7,666	crowd written paragraphs
fairy tales	C. Ovesdotter Alm	2005	15,000	sentences from children's stories
SemEval	C. Strapparava	2007	1,250	news headlines
TEC	S. M. Mohammad	2012	21,000	tweets
CBET	A. Gholipour Shahraki	2015	81,163	tweets

6.1.7. Chinese Lexicon. Lei et al. [2014] propose a framework of generating a domain- and context-dependent emotion lexicon. First, they select a well-formed training set from the corpus of news headlines taken from the Sina website, a popular news site in China. The criterion for selecting a headline is to be among those with the highest rating for at least one emotion. Next, the lexicon is built such that for each word f_j and each emotion e_k:

$$P(e_k|f_j) = \frac{\sum_{i=1}^{D} \sigma_{ij} r_{ik} \epsilon_i}{\sum_{k=1}^{E} \sum_{i=1}^{D} \sigma_{ij} r_{ik} \epsilon_i}, \tag{4}$$

where σ_{ij} is the relative term frequency of f_j in document d_i, r_{ik} is the co-occurrence number of document d_i and emotion e_k, and ϵ_i is the prior probability of document d_i. Results of their experiments show an improvement over existing lexicon generation methods such as in Katz et al. [2007].

6.2. Datasets

One of the old challenges in most machine-learning works is collecting data, especially labeled data. Apart from the costs of manual labeling, in the specific problem of emotion annotation, results are often subject to misunderstandings, subjective interpretations of annotators, their personality, the perspective that the content is analyzed, and so on [Alm 2008]. In this section, we introduce some useful datasets that have a reliable labeling process and/or are widely used. Table VIII shows a summary of these datasets.

6.2.1. ISEAR. Scherer and Wallbott [1994] present one of the oldest emotion-labeled datasets, ISEAR, which is freely available for download.[17] The data were collected during the 1990s by a large group of psychologists all over the world, who were working on the ISEAR project. In this survey, 3,000 students, including both psychologists and non-psychologists, from 37 countries on all five continents were asked to report situations in which they had experienced the following seven major emotions: *joy, fear, anger, sadness, disgust, shame*, and *guilt*. In what they write, respondents should explain how they had appraised the situation and how they reacted. For non-English speakers, the text was translated to English. Hence, the format of the data is a sentence or paragraph, labeled with exactly one emotion. This dataset is reliable in terms of labeling, since the authors themselves have annotated their text. However, translating from other languages to English might change the sense and emotions. Surprisingly, ISEAR was not used for emotion-mining purposes until 2008.

6.2.2. Fairy Tales. A set of fairy tales is another dataset[18] developed by Alm and Sproat [2005]. It contains 185 children's stories written by Beatrix Potter, Brothers Grimm, and Hans Christian Andersen, with a total of about 15,000 sentences that are labeled by one of the following emotions: *anger*, *disgust*, *fear*, *happiness*, *sadness*, *positively surprised*, *negatively surprised*, or *neutral* if it does not show any emotion. The annotation was

---

[17]http://www.affective-sciences.org/researchmaterial.
[18]http://people.rc.rit.edu/~coagla/affectdata/index.html.

done manually by six female native English speakers. Note that, unlike the ISEAR dataset, in which texts are annotated on the document level, in the fairy tales dataset, annotation is done on the sentence level.

6.2.3. SemEval 2007. Strapparava and Mihalcea [2007] developed a dataset for the SemEval 2007 workshop on the shared task of affective computing.[19] It consists of news headlines from major newspapers such as *The New York Times*, CNN, and BBC News, as well as the Google News search engine. The annotation was done manually by six annotators, and the set of labels includes six emotions: *anger*, *disgust*, *fear*, *joy*, *sadness*, and *surprise*. Instead of the usual 0/1 binary annotation, they run a finer-grained labeling process. An interval $[0, 100]$ is set for each emotion and the annotator decides to what degree from 0 to 100 the headline shows that emotion. Hence, a headline can have multiple emotions, each with a different degree. To justify why news are selected to build this dataset, they claim that news have typically a high load of emotional content and are written in a style meant to attract readers' attention. In fact, there is a popular concept in the news world, called "Emotional Framing" [Corcoran 2006], positing that each news item is shaped to a form of story with layers of dramatic frames, such as *fear* caused by danger or alarming news. Although this idea backs up the development of the SemEval 2007 dataset, our statistical analyses show that the data are most likely to be *neutral* and there is not much tangible emotion expressed by news. For example, the average degree of all emotions for a headline is only 15.48 (of 100) on average. Also, only 6.8%, 3.6%, 11.6%, 13.6%, 15.6%, and 3.2% of headlines express *anger*, *disgust*, *fear*, *joy*, *sadness*, and *surprise*, respectively, with a degree more than 50.

6.2.4. TEC. Mohammad [2012] created a corpus of emotional tweets from Twitter(TEC)[20] in 2012. He targeted the following six basic emotions proposed by Ekman et al. [1972]: *anger*, *disgust*, *fear*, *joy*, *sadness*, and *surprise*, and chose six hashtags addressing these emotions (e.g., #anger, #disgust, etc.) to search for appropriate tweets using Twitter Search Application Program Interface (API).[21] He discarded very short tweets, very badly spelled ones, and those with the prefix "RT," which are retweets of another tweet. He also removed the tweets that did not have the emotional hashtag at the end of the message, since he believed such hashtags may not be good indicators of the label of the tweet. After this post-processing, TEC includes 21,051 tweets where 7.4%, 3.6%, 13.4%, 39.1%, 18.2%, and 18.3% of the corpus have the labels *anger*, *disgust*, *fear*, *joy*, *sadness*, and *surprise*, respectively. This shows how imbalanced this dataset is.

6.2.5. CBET. In 2015, Gholipour Shahraki [2015] compiled the Cleaned Balanced Emotional Tweets (CBET) dataset[22] from Twitter using hashtags to search for tweets that have at least one of these nine emotions: *anger*, *fear*, *joy*, *love*, *sadness*, *surprise*, *thankfulness*, *disgust*, and *guilt*. The corpus is also preprocessed and cleaned. One interesting point in cleaning tweets exploited here is segmenting space-free phrases used as haghtags. For instance, the hashtag "#animalrights" is segmented to "animal" and "rights" while the original form of the hashtag is preserved as well. CBET has two parts: The larger part contains tweets that have exactly one label, referred to as single-labeled samples. This part is perfectly balanced over labels, containing 76,860 tweets, with 8,540 for each emotion. The smaller part contains double-labeled tweets, that is, those that express two emotions simultaneously. The size of this portion is 4,303 and it

---

is imbalanced, as not all combinations of emotions happen together equally frequently. The most frequent paired label is *joy-love*, while some pairs, such as *anger-thankfulness*, are very rare. The total number of 81,163 tweets in this dataset makes it the largest available corpus for emotion-mining research.

7. CONCLUSION

In this survey, we introduced state-of-the-art methods and improvements on text sentiment analysis. Sentiment analysis refers to all the areas of detecting, analyzing, and evaluating humans' state of mind towards different topics of interest. In particular, text sentiment analysis aims to mine people's opinions, sentiments, and emotions based on their writings. Personal notes, emails, news headlines, blogs, tales, novels, chat messages, and social networking websites such as Twitter, Facebook, and MySpace are some types of text that can convey emotions.

In this work, we suggested a careful categorization of tasks in this area and provided a clear and logical taxonomy of sentiment analysis work. There are two main subcategories in this field: opinion mining and emotion mining. The first one deals with the expression of opinions and the latter is concerned with the articulation of emotions. There is a rich body of research on opinion mining, and many new focused and specialized areas are investigated, while emotion mining from text is still in its infancy. Considering this fact and the strong link between them, we tried to give a comprehensive overview of the most recent trends and useful resources in opinion mining and emotion mining. Towards this goal, we first explained the key elements of the polarity classification task and reviewed those works in this area that can be useful for the emotion-mining task. Second, we introduced a set of important resources, including lexicons and datasets that researchers need for a polarity classification task. Third, we reviewed emotion theories as an introductory to the world of human emotions. A thorough survey on emotion-related research was given next and useful resources specific to emotion-mining work were introduced.

REFERENCES

Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. 2005. Emotions from text: Machine learning for text-based emotion prediction. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 579–586.

Cecilia Ovesdotter Alm and Richard Sproat. 2005. Emotional sequencing and development in fairy tales. In *Affective Computing and Intelligent Interaction*. Springer, 668–674.

Ebba Cecilia Ovesdotter Alm. 2008. *Affect in Text and Speech*. ProQuest.

Alina Andreevskaia and Sabine Bergler. 2006. Mining wordnet for a fuzzy sentiment: Sentiment tag extraction from wordnet glosses. In *EACL*, Vol. 6. 209–215.

Anthony Aue and Michael Gamon. 2005. Customizing sentiment classifiers to new domains: A case study. In *Proceedings of Recent Advances in Natural Language Processing (RANLP)*, Vol. 1. Citeseer.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *LREC*, Vol. 10. 2200–2204.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The berkeley framenet project. In *Proceedings of the 17th International Conference on Computational Linguistics-Volume 1*. Association for Computational Linguistics, 86–90.

Farah Benamara, Carmine Cesarano, Antonio Picariello, Diego Reforgiato Recupero, and Venkatramana S. Subrahmanian. 2007. Sentiment analysis: Adjectives and adverbs are better than adjectives alone. In *ICWSM*.

Plaban Kumar Bhowmick. 2009. Reader perspective emotion analysis in text through ensemble based multi-label classification framework. *Comput. Inf. Sci.* 2, 4 (2009), 64–74.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.* 3 (2003), 993–1022.

John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *ACL*, Vol. 7. Citeseer, 440–447.

Johan Bollen, Huina Mao, and Alberto Pepe. 2011. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In *ICWSM*.

Carlos Busso, Zhigang Deng, Serdar Yildirim, Murtaza Bulut, Chul Min Lee, Abe Kazemzadeh, Sungbok Lee, Ulrich Neumann, and Shrikanth Narayanan. 2004. Analysis of emotion recognition using facial expressions, speech and multimodal information. In *Proceedings of the 6th International Conference on Multimodal Interfaces*. ACM, 205–211.

François-Régis Chaumartin. 2007. UPAR7: A knowledge-based system for headline sentiment tagging. In *Proceedings of the 4th International Workshop on Semantic Evaluations*. Association for Computational Linguistics, 422–425.

P. E. Corcoran. 2006. Emotional framing in australian journalism. In *Australian & New Zealand Communication Association International Conference, Adelaide, Australia (ANZCA)*.

Taner Danisman and Adil Alpkocak. 2008. Feeler: Emotion classification of text using vector space model. In *AISB 2008 Convention Communication, Interaction and Social Intelligence*, Vol. 1. 53.

Amitava Das and Sivaji Bandyopadhyay. 2010. SentiWordNet for indian languages. *Asian Federation for Natural Language Processing, China* (2010), 56–63.

Sanjiv Das and Mike Chen. 2001. Yahoo! for amazon: Extracting market sentiment from stock message boards. In *Proceedings of the Asia Pacific Finance Association Annual Conference (APFA)*, Vol. 35. Bangkok, Thailand, 43.

Kushal Dave, Steve Lawrence, and David M. Pennock. 2003. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th International Conference on World Wide Web*. ACM, 519–528.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. In *ICWSM*.

Kerstin Denecke and Yihan Deng. 2015. Sentiment analysis in medical settings: New opportunities and challenges. *Artif. Intell. Med.* 64, 1 (2015), 17–27.

Sidney K. D'mello and Jacqueline Kory. 2015. A review and meta-analysis of multimodal affect detection systems. *ACM Comput. Surv.* 47, 3 (2015), 43.

Cícero Nogueira dos Santos and Maira Gatti. 2014. Deep convolutional neural networks for sentiment analysis of short texts. In *COLING*. 69–78.

Paul Ekman. 1992. An argument for basic emotions. *Cogn. Emot.* 6, 3–4 (1992), 169–200.

Paul Ekman, Wallace V. Friesen, and Phoebe Ellsworth. 1972. Emotion in the human face: Guidelines for research and an integration of findings. New York. Permagon.

Moataz El Ayadi, Mohamed S. Kamel, and Fakhri Karray. 2011. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recogn.* 44, 3 (2011), 572–587.

Andrea Esuli and Fabrizio Sebastiani. 2005. Determining the semantic orientation of terms through gloss classification. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*. ACM, 617–624.

Andrea Esuli and Fabrizio Sebastiani. 2006a. Determining term subjectivity and term orientation for opinion mining. In *EACL*, Vol. 6. 2006.

Andrea Esuli and Fabrizio Sebastiani. 2006b. Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of LREC*, Vol. 6. 417–422.

Aidan Finn and Nicholas Kushmerick. 2006. Learning to classify documents according to genre. *J. Am. Soc. Inf. Sci. Technol.* 57, 11 (2006), 1506–1518.

Elaine Fox. 2008. *Emotion Science Cognitive and Neuroscientific Approaches to Understanding Human Emotions*. Palgrave Macmillan.

Michael Gamon. 2004. Sentiment classification on customer feedback data: Noisy data, large feature vectors, and the role of linguistic analysis. In *Proceedings of the 20th International Conference on Computational Linguistics*. Association for Computational Linguistics, 841.

Michael Gamon and Anthony Aue. 2005. Automatic identification of sentiment vocabulary: Exploiting low association with known sentiment terms. In *Proceedings of the ACL Workshop on Feature Engineering for Machine Learning in Natural Language Processing*. Association for Computational Linguistics, 57–64.

Bo Gao, Bettina Berendt, and Joaquin Vanschoren. 2015a. Who is more positive in private? Analyzing sentiment differences across privacy levels and demographic factors in facebook chats and posts. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*. ACM, 605–610.

Kai Gao, Hua Xu, and Jiushuo Wang. 2015b. Emotion cause detection for chinese micro-blogs based on ECOCC model. In *Advances in Knowledge Discovery and Data Mining*. Springer, 3–14.

Ameneh Gholipour Shahraki. 2015. *Emotion Detection from Text*. Master's thesis. University of Alberta.

Narendra Gupta, Mazin Gilbert, and Giuseppe Di Fabbrizio. 2013. Emotion detection in email customer care. *Comput. Intell.* 29, 3 (2013), 489–505.

Jeffrey T. Hancock, Christopher Landrigan, and Courtney Silver. 2007. Expressing emotion in text-based communication. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 929–932.

Maryam Hasan, Emmanuel Agu, and Elke Rundensteiner. 2014. Using hashtags as labels for supervised learning of emotions in Twitter messages. In *Proceedings of the Health Informatics Workshop (HI-KDD)*.

Vasileios Hatzivassiloglou and Kathleen R. McKeown. 1997. Predicting the semantic orientation of adjectives. In *Proceedings of the 8th Conference on European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 174–181.

Vasileios Hatzivassiloglou and Janyce M. Wiebe. 2000. Effects of adjective orientation and gradability on sentence subjectivity. In *Proceedings of the 18th Conference on Computational Linguistics-Volume 1*. Association for Computational Linguistics, 299–305.

Dung T. Ho and Tru H. Cao. 2012. A high-order hidden Markov model for emotion detection from textual data. In *Knowledge Management and Acquisition for Intelligent Systems*. Springer, 94–105.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 168–177.

Xia Hu, Jiliang Tang, Huiji Gao, and Huan Liu. 2013. Unsupervised sentiment analysis with emotional signals. In *Proceedings of the 22nd International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 607–618.

Mukesh Jain and V. Kulkarni. 2014. TexEmo: Conveying emotion from text-the study. *Int. J. Comput. Appl.* 86, 4 (2014), 43–49.

Jing Jiang and ChengXiang Zhai. 2007. Instance weighting for domain adaptation in NLP. In *ACL*, Vol. 7. Citeseer, 264–271.

Nitin Jindal and Bing Liu. 2008. Opinion spam and analysis. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*. ACM, 219–230.

Yohan Jo and Alice H. Oh. 2011. Aspect and sentiment unification model for online review analysis. In *Proceedings of the 4th ACM International Conference on Web Search and Data Mining*. ACM, 815–824.

Nobuhiro Kaji and Masaru Kitsuregawa. 2006. Automatic construction of polarity-tagged corpus from HTML documents. In *Proceedings of the COLING/ACL on Main Conference Poster Sessions*. Association for Computational Linguistics, 452–459.

Nobuhiro Kaji and Masaru Kitsuregawa. 2007. Building lexicon for sentiment analysis from massive collection of HTML documents.. In *EMNLP-CoNLL*. Citeseer, 1075–1083.

Jaap Kamps, M. J. Marx, Robert J. Mokken, and Maarten De Rijke. 2004. Using wordnet to measure semantic orientations of adjectives. *Language Resources and Evaluation Conference (LREC)* 4 (2004), 1115–1118.

Daekook Kang and Yongtae Park. 2014. Review-based measurement of customer satisfaction in mobile service: Sentiment analysis and VIKOR approach. *Expert Syst. Appl.* 41, 4 (2014), 1041–1050.

E. C.-C. Kao, Chun-Chieh Liu, Ting-Hao Yang, Chang-Tai Hsieh, and Von-Wun Soo. 2009. Towards text-based emotion detection a survey and possible improvements. In *Proceedings of the 2009 International Conference on Information Management and Engineering (ICIME'09)*. IEEE, 70–74.

Phil Katz, Matthew Singleton, and Richard Wicentowski. 2007. Swat-mp: The semeval-2007 systems for task 5 and task 14. In *Proceedings of the 4th International Workshop on Semantic Evaluations*. Association for Computational Linguistics, 308–313.

Sunghwan Mac Kim, Alessandro Valitutti, and Rafael A Calvo. 2010. Evaluation of unsupervised emotion models to textual affect recognition. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*. Association for Computational Linguistics, 62–70.

Svetlana Kiritchenko, Xiaodan Zhu, Colin Cherry, and Saif Mohammad. 2014. NRC-Canada-2014: Detecting aspects and sentiment in customer reviews. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*. Association for Computational Linguistics and Dublin City University, Dublin, Ireland, 437–442.

Andrea Kleinsmith and Nadia Bianchi-Berthouze. 2013. Affective body expression perception and recognition: A survey. *IEEE Trans. Affect. Comput.* 4, 1 (2013), 15–33.

Sophia Yat Mei Lee, Ying Chen, Shoushan Li, and Chu-Ren Huang. 2010. Emotion cause events: Corpus construction and analysis. In *LREC*.

Jingsheng Lei, Yanghui Rao, Qing Li, Xiaojun Quan, and Liu Wenyin. 2014. Towards building a social emotion detection system for online news. *Fut. Gen. Comput. Syst.* 37 (2014), 438–448.

Shoushan Li, Yunxia Xue, Zhongqing Wang, and Guodong Zhou. 2013. Active learning for cross-domain sentiment classification. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence*. AAAI Press, 2127–2133.

Tao Li, Vikas Sindhwani, Chris Ding, and Yi Zhang. 2009. Knowledge transformation for cross-domain sentiment classification. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 716–717.

Weiyuan Li and Hua Xu. 2014. Text-based emotion classification using emotion cause extraction. *Expert Syst. Appl.* 41, 4 (2014), 1742–1749.

Chenghua Lin and Yulan He. 2009. Joint sentiment/topic model for sentiment analysis. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*. ACM, 375–384.

Wei-Hao Lin, Theresa Wilson, Janyce Wiebe, and Alexander Hauptmann. 2006. Which side are you on?: Identifying perspectives at the document and sentence levels. In *Proceedings of the 10th Conference on Computational Natural Language Learning*. Association for Computational Linguistics, 109–116.

Bing Liu. 2011. Opinion mining and sentiment analysis. In *Web Data Mining*. Springer, 459–526.

Bing Liu. 2012. Sentiment analysis and opinion mining. *Synth. Lect. Hum. Lang. Technol.* 5, 1 (2012), 1–167.

Bing Liu. 2015. *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. Cambridge University Press.

Bing Liu, Minqing Hu, and Junsheng Cheng. 2005. Opinion observer: Analyzing and comparing opinions on the web. In *Proceedings of the 14th International Conference on World Wide Web*. ACM, 342–351.

Hugo Lövheim. 2012. A new three-dimensional model for emotions and monoamine neurotransmitters. *Med. Hypoth.* 78, 2 (2012), 341–348.

Kim Luyckx, Frederik Vaassen, Claudia Peersman, and Walter Daelemans. 2012. Fine-grained emotion detection in suicide notes: A thresholding approach to multi-label classification. *Biomed. Inf. Insights* 5, Suppl. 1 (2012), 61.

D. A. Medler, A. Arnoldussen, J. R. Binder, and M. S. Seidenberg. 2005. The Wisconsin Perceptual Attribute Ratings Database. Retrieved from http://www.neuro.mcw.edu/ratings/.

Prem Melville, Wojciech Gryc, and Richard D. Lawrence. 2009. Sentiment analysis of blogs by combining lexical knowledge with text classification. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1275–1284.

Gilad Mishne. 2005. Experiments with mood classification in blog posts. In *Proceedings of ACM SIGIR 2005 Workshop on Stylistic Analysis of Text for Information Access*, Vol. 19. Citeseer.

Gilad Mishne and Maarten De Rijke. 2006. Capturing global mood levels using blog posts. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*. 145–152.

Saif M. Mohammad and Peter D. Turney. 2010. Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*. Association for Computational Linguistics, 26–34.

Saif M. Mohammad. 2012. # Emotional tweets. In *Proceedings of the 1st Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the 6th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, 246–255.

Saif M. Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets. In *Proceedings of the 7th International Workshop on Semantic Evaluation Exercises (SemEval'13)*.

Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi. 2011. Towards multimodal sentiment analysis: Harvesting opinions from the web. In *Proceedings of the 13th International Conference on Multimodal Interfaces*. ACM, 169–176.

Arjun Mukherjee and Bing Liu. 2012. Aspect extraction through semi-supervised modeling. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*. Association for Computational Linguistics, 339–348.

Tony Mullen and Nigel Collier. 2004. Sentiment analysis using support vector machines with diverse information sources. In *EMNLP*, Vol. 4. 412–418.

Sean M. Murphy, Bernard Maskit, and Wilma Bucci. 2015. Putting feelings into words: Cross-linguistic markers of the referential process. *NAACL HLT 2015* (2015), 80.

Jin-Cheon Na, Haiyang Sui, Christopher Khoo, Syin Chan, and Yunyun Zhou. 2004. Effectiveness of simple linguistic processing in automatic sentiment classification of product reviews. *Adv. Knowl. Org.* 9 (2004), 49–54.

Tetsuya Nasukawa and Jeonghee Yi. 2003. Sentiment analysis: Capturing favorability using natural language processing. In *Proceedings of the 2nd International Conference on Knowledge Capture*. ACM, 70–77.

Alena Neviarouskaya, Helmut Prendinger, and Mitsuru Ishizuka. 2007. Textual affect sensing for sociable and expressive online communication. In *Affective Computing and Intelligent Interaction*. Springer, 218–229.

Alena Neviarouskaya, Helmut Prendinger, and Mitsuru Ishizuka. 2009. Compositionality principle in recognition of fine-grained emotions from text. In *ICWSM*.

Vincent Ng, Sajib Dasgupta, and S. M. Arifin. 2006. Examining the role of linguistic knowledge sources in the automatic identification and classification of reviews. In *Proceedings of the COLING/ACL on Main Conference Poster Sessions*. Association for Computational Linguistics, 611–618.

Alvaro Ortigosa, José M. Martín, and Rosa M. Carro. 2014. Sentiment analysis in facebook and its application to e-learning. *Comput. Hum. Behav.* 31 (2014), 527–541.

Sinno Jialin Pan, Xiaochuan Ni, Jian-Tao Sun, Qiang Yang, and Zheng Chen. 2010. Cross-domain sentiment classification via spectral feature alignment. In *Proceedings of the 19th International Conference on World Wide Web*. ACM, 751–760.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 271.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 115–124.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Found. Trends Inf. Retriev.* 2, 1–2 (2008), 1–135.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: Sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing-Volume 10*. Association for Computational Linguistics, 79–86.

James W. Pennebaker. 1997. Writing about emotional experiences as a therapeutic process. *Psychol. Sci.* 8, 3 (1997), 162–166.

James W. Pennebaker, R. J. Booth, and M. E. Francis. 2007. Linguistic inquiry and word count: LIWC. *Austin, TX: liwc. net* (2007).

Robert Plutchik and Henry Kellerman. 1986. *Emotion: Theory, Research and Experience*. Vol. 3. Academic Press, New York, NY.

Rudy Prabowo and Mike Thelwall. 2009. Sentiment analysis: A combined approach. *J. Inform.* 3, 2 (2009), 143–157.

Likun Qiu, Weishi Zhang, Changjian Hu, and Kai Zhao. 2009. Selc: A self-supervised model for sentiment classification. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*. ACM, 929–936.

Francisco Rangel and Paolo Rosso. 2016. On the impact of emotions on author profiling. *Information Processing & Management* 52, 1 (2016), 73–92.

Yanghui Rao, Qing Li, Liu Wenyin, Qingyuan Wu, and Xiaojun Quan. 2014. Affective topic model for social emotion detection. *Neur. Netw.* 58 (2014), 29–37.

Kirk Roberts, Michael A. Roach, Joseph Johnson, Josh Guthrie, and Sanda M. Harabagiu. 2012. EmpaTweet: Annotating and detecting emotions on Twitter. In *LREC*. 3806–3813.

James A. Russell, José-Miguel Fernández-Dols, Anthony S. R. Manstead, and Jane C. Wellenkamp. 2013. *Everyday Conceptions of Emotion: An Introduction to the Psychology, Anthropology and Linguistics of Emotion*. Vol. 81. Springer Science & Business Media.

Robert E. Schapire. 1999. A brief introduction to boosting. In *Ijcai*, Vol. 99. 1401–1406.

Klaus R. Scherer and Harald G. Wallbott. 1994. Evidence for universality and cultural variation of differential emotion response patterning. *J. Pers. Soc. Psychol.* 66, 2 (1994), 310.

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

Phillip Shaver, Judith Schwartz, Donald Kirson, and Cary O'connor. 1987. Emotion knowledge: Further exploration of a prototype approach. *J. Pers. Soc. Psychol.* 52, 6 (1987), 1061.

Richard Socher, Alex Perelygin, Jean Y. Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Vol. 1631. Citeseer, 1642.

Philip Stone, Dexter C. Dunphy, Marshall S. Smith, and D. M. Ogilvie. 1968. The general inquirer: A computer approach to content analysis. *J. Region. Sci.* 8, 1 (1968), 113–116.

Carlo Strapparava and Rada Mihalcea. 2007. Semeval-2007 task 14: Affective text. In *Proceedings of the 4th International Workshop on Semantic Evaluations*. Association for Computational Linguistics, 70–74.

Carlo Strapparava and Rada Mihalcea. 2008. Learning to identify emotions in text. In *Proceedings of the 2008 ACM Symposium on Applied Computing*. ACM, 1556–1560.

Carlo Strapparava and Alessandro Valitutti. 2004. WordNet affect: An affective extension of wordnet. In *LREC*, Vol. 4. 1083–1086.

Hiroya Takamura, Takashi Inui, and Manabu Okumura. 2007. Extracting semantic orientations of phrases from dictionary. In *HLT-NAACL*, Vol. 2007. 292–299.

Songbo Tan, Gaowei Wu, Huifeng Tang, and Xueqi Cheng. 2007. A novel scheme for domain-transfer problem in the context of sentiment analysis. In *Proceedings of the 16th ACM Conference on Conference on Information and Knowledge Management*. ACM, 979–982.

Duyu Tang. 2015. Sentiment-specific representation learning for document-level sentiment analysis. In *Proceedings of the 8th ACM International Conference on Web Search and Data Mining*. ACM, 447–452.

Duyu Tang, Furu Wei, Bing Qin, Ting Liu, and Ming Zhou. 2014a. Coooolll: A deep learning system for Twitter sentiment classification. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*. 208–212.

Duyu Tang, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, and Bing Qin. 2014b. Learning sentiment-specific word embedding for Twitter sentiment classification. In *ACL (1)*. 1555–1565.

Peter D. Turney. 2002. 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*. Association for Computational Linguistics, 417–424.

S. Voeffray. 2011. Emotion-sensitive human-computer interaction (HCI): State of the art-Seminar paper. *Emotion Recognition. p1-4* (2011).

Xuan-Son Vu and Seong-Bae Park. 2014. Construction of vietnamese SentiWordNet by using vietnamese dictionary. In *Proceedings of the 40th Conference of the Korea Information Processing Society*. 745–748.

Joseph B. Walther. 1992. Interpersonal effects in computer-mediated interaction a relational perspective. *Commun. Res.* 19, 1 (1992), 52–90.

Joseph B. Walther, Tracy Loh, and Laura Granka. 2005. Let me count the ways the interchange of verbal and nonverbal cues in computer-mediated and face-to-face affinity. *J. Lang. Soc. Psychol.* 24, 1 (2005), 36–65.

Shuai Wang, Zhiyuan Chen, and Bing Liu. 2016. Mining aspect-specific opinion using a holistic lifelong topic model. In *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 167–176.

Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan, and Amit P. Sheth. 2012. Harnessing Twitter "big data" for automatic emotion identification. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*. IEEE, 587–592.

Casey Whitelaw, Navendu Garg, and Shlomo Argamon. 2005. Using appraisal groups for sentiment analysis. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*. ACM, 625–631.

Janyce Wiebe, Theresa Wilson, Rebecca Bruce, Matthew Bell, and Melanie Martin. 2004. Learning subjective language. *Comput. Ling.* 30, 3 (2004), 277–308.

Alicja Wieczorkowska, Piotr Synak, and Zbigniew W. Raś. 2006. Multi-label classification of emotions in music. In *Intelligent Information Processing and Web Mining*. Springer, 307–315.

Theresa Wilson, Paul Hoffmann, Swapna Somasundaran, Jason Kessler, Janyce Wiebe, Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. 2005a. OpinionFinder: A system for subjectivity analysis. In *Proceedings of hlt/emnlp on Interactive Demonstrations*. Association for Computational Linguistics, 34–35.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005b. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 347–354.

Hui Yang, Jamie Callan, and Luo Si. 2006. Knowledge transfer and opinion detection in the TREC 2006 blog track. In *TREC*.

Yi-Hsuan Yang and Homer H. Chen. 2012. Machine recognition of music emotion: A review. *ACM Trans. Intell. Syst. Technol.* 3, 3 (2012), 40.

Hong Yu and Vasileios Hatzivassiloglou. 2003. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 129–136.

Zhihong Zeng, Maja Pantic, Glenn I. Roisman, and Thomas S. Huang. 2009. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Trans. Pattern Anal. Mach. Intell.* 31, 1 (2009), 39–58.

Yaowen Zhang, Lin Shang, and Xiuyi Jia. 2015. Sentiment analysis on microblogging by integrating text and image features. In *Advances in Knowledge Discovery and Data Mining*. Springer, 52–63.

Xiaodan Zhu, Svetlana Kiritchenko, and Saif M. Mohammad. 2014. NRC-canada-2014: Recent improvements in the sentiment analysis of tweets. *SemEval 2014* (2014), 443.

Xiaodan Zhu, Svetlana Kiritchenko, and Saif M. Mohammad. 2014. Sentiment analysis of short informal texts. *J. Artif. Intell. Res.* 50 (2014), 723–762.