Submitted to the Special issue on Semantic Analysis in Social Media, Computational Intelligence.
Guest editors: Atefeh Farzindar (farzindarATnlptechnologiesDOTca), Diana Inkpen (dianaATeecsDOTuottawaDOTca)

Using Hashtags to Capture Fine Emotion Categories from Tweets

Saif M. Mohammad and Svetlana Kiritchenko
National Research Council Canada.
Ottawa, Ontario, Canada, K1A 0R6
{saif.mohammad,svetlana.kiritchenko}@nrc-cnrc.gc.ca

Detecting emotions in microblogs and social media posts has applications for industry, health, and security. Statistical, supervised automatic methods for emotion detection rely on text that is labeled for emotions, but such data is rare and available for only a handful of basic emotions. In this paper, we show that emotion-word hashtags are good manual labels of emotions in tweets. We also propose a method to generate a large lexicon of word–emotion associations from this emotion-labeled tweet corpus. This is the first lexicon with real-valued word–emotion association scores. We begin with experiments for six basic emotions and show that the hashtag annotations are consistent and match with the annotations of trained judges. We also show how the extracted tweets corpus and word–emotion associations can be used to improve emotion classification accuracy in a different non-tweets domain.

Eminent psychologist, Robert Plutchik, had proposed that emotions have a relationship with personality traits. However, empirical experiments to establish this relationship have been stymied by the lack of comprehensive emotion resources. Since personality may be associated with any of the hundreds of emotions, and since our hashtag approach scales easily to a large number of emotions, we extend our corpus by collecting tweets with hashtags pertaining to 585 fine emotions. Then, for the first time, we present experiments to show that fine emotion categories such as that of excitement, guilt, yearning, and admiration are useful in automatically detecting personality from text. Stream-of-consciousness essays and collections of Facebook posts marked with personality traits of the author are used as the test sets.

Key words: Emotions, affect, tweets, social media, hashtags, basic emotions, personality detection, Big 5 model, word–emotion associations, sentiment analysis.

1. INTRODUCTION

We use language not just to convey facts, but also our emotions. For example, given the sentence, *That jerk stole my photo on tumblr*, it is easy to deduce that the speaker is angry. Clues to emotion are often present at a lexical level. For example, *delightful* and *yummy* indicate the emotion of joy, *gloomy* and *cry* indicate sadness, *shout* and *jerk* indicate anger, and so on. Automatically identifying emotions expressed in text has a number of applications, including customer relation management (Bougie *et al.*, 2003), determining popularity of products and governments (Mohammad and Yang, 2011), identifying high-risk suicide cases (Osgood and Walker, 1959; Matykiewicz *et al.*, 2009; Pestian *et al.*, 2008; Cherry *et al.*, 2012), improving human-computer interaction (Velásquez, 1997; Ravaja *et al.*, 2006), and automatic tutoring systems (Litman and Forbes-Riley, 2004).

Most statistical automatic approaches are supervised and require large amounts of labeled data. Manual annotation of text with emotions is time-intensive and costly. Thus only a small amount of such text exists. Examples include 1,250 newspaper headlines (Strapparava and Mihalcea, 2007) and about 4000 blog sentences (Aman and Szpakowicz, 2007) that are classified as expressing one of six basic emotions (joy, sadness, anger, fear, disgust, and surprise) or neutral. However, humans are capable of distinguishing and expressing a few hundred different emotions such as guilt, remorse, optimism, and enthusiasm (not just six). As we will show through experiments, identifying these fine-grained emotions are useful in applications such as personality detection.

None of the existing datasets contain emotion-annotated text from social media websites such as Facebook and Twitter. Twitter is an online social networking and microblogging service where users post and read messages that are up to 140 characters long. Unlike Facebook posts, which are largely private and often visible only to friends of the poster, Twitter posts are public and visible to all. The Twitter posts are called *tweets*. The people who post these messages are called *tweeters*. Often a tweet may include one or more words immediately preceded with a hash symbol (#). These

words are called *hashtags*. Hashtags serve many purposes, but most notably they are used to indicate the topic. Often these words add to the information in the tweet: for example, hashtags indicating the tone of the message or the tweeter's emotions.

From the perspective of one consuming tweets, hashtags play a role in search: Twitter allows people to search tweets not only through words in the tweets, but also through hashtagged words. Consider the tweet below:

We are fighting for the 99% that have been left behind. #OWS #anger

A number of people tweeting about the Occupy Wall Street movement added the hashtag *#OWS* to their tweets. This allowed people searching for tweets about the movement to access them simply by searching for the *#OWS* hashtag. In this particular instance, the tweeter has also added an emotion-word hashtag *#anger*, possibly to convey that he or she is angry.

In this paper, we show that emotion word hashtags in tweets, such as *#angry* above, can be used as labeled data by a supervised learning system for emotion detection. Note that these emotion labels are not assigned by somebody other than the author (as is the case in traditional annotation), but rather these are emotion labels given by the tweeters themselves to their own messages, and corresponding to their own emotions at the time the message was composed.

We first create a large corpus of such emotion-labeled tweets for six basic emotions (Section 3). We will refer to this dataset as the Hashtag Emotion Corpus. We show how the Hashtag Emotion Corpus can be used for automatic emotion detection in tweets and also to improve automatic emotion detection in a different domain (newspaper headlines) (Section 4). We show through experiments that even though the tweets and hashtags cover a diverse array of topics and were generated by thousands of different individuals (possibly with very different educational and socio-economic backgrounds), the hashtag-based annotations of emotion are consistent and match the intuitions of trained judges.

A word–emotion association lexicon is a list of words and associated emotions. They can be used in numerous applications of emotion detection, such as personality detection, automatic dialogue systems, automatic tutoring systems, customer relation models, and even simply for highlighting words and phrases in a piece of text to quickly convey regions of affect. In Section 5, we describe how we extract a word–emotion association lexicon from the Hashtag Emotion Corpus. We will refer to it as the *Hashtag Emotion Lexicon*. In addition to word–emotion pairs, each entry in this lexicon also comes with a real-valued score indicating the degree of association between the word and the emotion. Higher scores indicate greater association. We first show how the word–emotion lexicon helps in emotion detection, and then use it to improve personality detection from text as described below.

Plutchik (1962) proposed that emotions have a relationship with personality. However, empirical experiments to establish the relationship have been stymied by the lack of comprehensive emotion resources—Humans are capable of feeling and expressing hundreds of different emotions. Since our hashtag approach can easily be scaled up to work with more emotions, we extend our corpus by collecting tweets with hashtags pertaining to 585 fine emotions (Section 6). We also expand the Hashtag Emotion Lexicon to include entries for each of the 585 emotions. This is the first lexicon with real-valued word–emotion association scores for hundreds of emotions. [1]

Then, for the first time, we present experiments to show that fine emotion categories such as that of excitement, guilt, yearning, and admiration are useful in automatically detecting personality from text (Section 7). Personality detection from text is the task of automatically detecting a person's personality traits, such as extroversion and agreeability, from free-form text written by her. As a testbed for our experiments we use standard personality-labeled texts used in prior work:

- *Stream-of-consciousness essays*: a collection of 2,469 essays (1.9 million words) and associated Big 5 personality traits compiled by Pennebaker and King (1999).

- *Collections of Facebook posts*: a collection of 10,000 Facebook posts (status updates) of 250 users and associated Big 5 personality traits compiled by Kosinski *et al.* (2013).

[1]Email Saif Mohammad to obtain a copy of the Hashtag Emotion Corpus or the Hashtag Emotion Lexicon: saif.mohammad@nrc-cnrc.gc.ca.

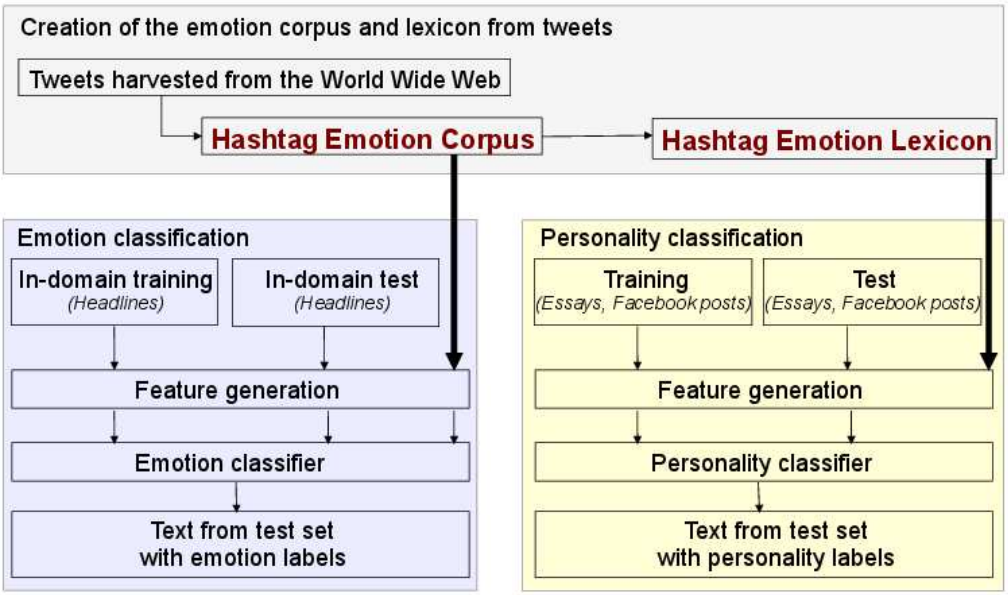

Figure 1. An overview of the resources created and the main classification experiments presented in this paper. Some of the experiments described in this paper are not shown here to avoid clutter.

Additionally, we perform experiments to show that the gains provided by the fine affect categories are not obtained by using coarse affect categories alone. Thus showing the benefit of modeling hundreds of affect categories, as opposed to modeling only a handful of basic emotions, in an extrinsic task. Figure 1 gives an overview of the resources created and the main classification experiments presented in this paper.

The paper is organized as follows: We begin with a description of related work (Section 2). Next we show how we create emotion-labeled text from tweets (Section 3). Section 4 presents experiments that demonstrate that the emotion labels are consistent and useful for automatic emotion detection. In Section 5, we show how we create a word–emotion lexicon from the tweets. In Section 6, we describe how we extend the labeled tweets corpus and also the word–emotion lexicon from a handful of emotions to a few hundred emotions. In Section 7, we present experiments to demonstrate that the Hashtag Emotion Lexicon is useful in the extrinsic task of detecting personality of authors from their essays and from collections of their Facebook posts. We conclude and present future research directions in Section 8.

2. RELATED WORK

We present below related work on emotion detection, past work on creating emotion labeled text, and some previous approaches to personality detection.

2.1. Emotion Detection

Emotion analysis can be applied to all kinds of text, but certain domains and modes of communication tend to have more overt expressions of emotions than others. Genereux and Evans (2006), Mihalcea and Liu (2006), and Neviarouskaya *et al.* (2009) analyzed web-logs. Alm *et al.* (2005) and Francisco and Gervás (2006) worked on fairy tales. Boucouvalas (2002), John *et al.* (2006), and Mohammad (2012a) explored emotions in novels. Zhe and Boucouvalas (2002), Holzman and Holzman and Pottenger (2003), and Ma *et al.* (2005) annotated chat messages for emotions. Liu

TABLE 1. Inter-annotator agreement (Pearson's correlation) amongst 6 annotators on the 1000-headlines dataset.

emotion	# of instances	% of instances	r
anger	132	13.2	0.50
disgust	43	4.3	0.45
fear	247	24.7	0.64
joy	344	34.4	0.60
sadness	283	28.3	0.68
surprise	253	25.3	0.36
		simple average	0.54
		frequency-based average	0.43

et al. (2003) and Mohammad and Yang (2011) worked on email data. Kim *et al.* (2009) analyzed sadness in posts reacting to news of Michael Jackson's death. Tumasjan *et al.* (2010) study Twitter as a forum for political deliberation.

Much of this work focuses on six emotions argued by Ekman (1992) to be the most basic emotions: joy, sadness, anger, fear, disgust, and surprise. Plutchik (1962, 1980, 1994) proposes a theory with eight basic emotions. These include Ekman's six as well as trust and anticipation. There is less work on other emotions, for example, work by Pearl and Steyvers (2010) that focuses on politeness, rudeness, embarrassment, formality, persuasion, deception, confidence, and disbelief. Bollen *et al.* (2011) measured tension, depression, anger, vigor, fatigue, and confusion in tweets. Francisco and Gervás (2006) marked sentences in fairy tales with tags for pleasantness, activation, and dominance, using lexicons of words associated with the three categories. One of the advantages of our work is that we can easily collect tweets with hashtags for hundreds of emotions, well beyond the basic six. This labeled data is a valuable resource in building automatic systems for detecting these emotions in text.

Go *et al.* (2009) and González-Ibáñez *et al.* (2011) noted that sometimes people use the hashtag *#sarcasm* to indicate that their tweet is sarcastic. They collected tweets with hashtags of *#sarcasm* and *#sarcastic* to create a dataset of sarcastic tweets. We follow their ideas and collect tweets with hashtags pertaining to different emotions. Additionally, we present several experiments to validate that the emotion labels in the corpus are consistent and match intuitions of trained judges.

2.2. Existing Emotion-Labeled Text

A number of text resources exist that are labeled for positive and negative sentiments, including movie and product reviews, newspaper headlines, blog posts, and tweets. However, text labeled with fine emotion labels is scarce. The SemEval-2007 Affective Text corpus has newspaper headlines labeled with the six Ekman emotions by six annotators (Strapparava and Mihalcea, 2007). More precisely, for each headline–emotion pair, the annotators gave scores from 0 to 100 indicating how strongly the headline expressed the emotion. The inter-annotator agreement as determined by calculating the Pearson's product moment correlation (r) between the scores given by each annotator and the average of the other five annotators is shown in Table 1. For our experiments, we considered scores greater than 25 to indicate that the headline expresses the corresponding emotion.

The dataset was created for an unsupervised competition, and consisted of 250 headlines of trial data and 1000 headlines of test data. We will refer to them as the 250-headlines and the 1000-headlines datasets respectively. However, the data has also been used in a supervised setting through (1) ten-fold cross-validation on the 1000-headlines dataset and (2) using the 1000 headlines as training data and testing on the 250-headlines dataset (Chaffar and Inkpen, 2011).

Other datasets with sentence-level annotations of emotions include about 4000 sentences from blogs, compiled by Aman and Szpakowicz (2007); 1000 sentences from stories on topics such as education and health, compiled by Neviarouskaya *et al.* (2009); and about 4000 sentences from fairy tales, annotated by Alm and Sproat (2005).

2.3. Personality Detection

Personality has significant impact on our lives—for example, on job performance (Tett *et al.*, 1991), inter-personal relations (White *et al.*, 2004), on the products we buy (Lastovicka and Joachimsthaler, 1988), and even on our health and well-being (Hayes and Joseph, 2003). The five-factor or the big five model of personality describes personality along the dimensions listed below:

- extroversion vs. introversion: sociable, assertive vs. aloof, shy
- neuroticism vs. emotional stability: insecure, anxious vs. calm, unemotional
- agreeability vs. disagreeability: friendly, cooperative vs. antagonistic, fault-finding
- conscientiousness vs. unconscientiousness: self-disciplined, organized vs. inefficient, careless
- openness to experience vs. conventionality: intellectual, insightful vs. shallow, unimaginative

There exist other models of personality too, such as the Myers-Brigg Type Indicator (Myers, 1962; Myers *et al.*, 1985). However, the big five model has been accepted more widely by the research community. The five dimensions listed above have been shown to entail various other personality traits.

Traditionally, personality is determined through specific questionnaires. However, automatically identifying personality from free-form text is far more desirable. Some of the earliest work on automatic personality detection was by Pennebaker and King (1999). They asked students to write stream-of-consciousness essays, that is, unedited pieces of text written in one sitting, detailing what was in their mind without conscious effort to structure their thoughts. The students were later provided with questionnaires to assess their Big 5 personality traits. Pennebaker and King (1999) used lexical categories from Linguistic Inquiry and Word Count (LIWC) to identify linguistic correlates of personality.[2] They showed, for example, that agreeability is characterized with more positive emotion words and fewer articles and that neurotism is characterized with more negative emotion words and more first-person pronouns. Openness to experience is correlated with longer words and avoidance of first-person pronouns, and conscientiousness with fewer negations and negative words. Mairesse *et al.* (2007) improved on these features and distribute their system online.[3] They also use features such as imageability from the MRC Psycholinguistic Database Machine Usable Dictionary.[4] Both Pennebaker and King (1999) and Mairesse *et al.* (2007) worked with the Essays dataset. More recently, there has also been work on personality detection from blogs (Yarkoni, 2010), collections of Facebook posts (Kosinski *et al.*, 2013), and collections of Twitter posts and follower network (Qiu *et al.*, 2012). There also exist websites that analyze blogs and display the personality types of the authors.[5]

Plutchik (1962) states that persistent situations involving emotions produce persistent traits or personality. For example, if one is angry most of the time, then anger or related phenomenon such as aggressiveness become part of the personality. Emotions are considered to be more transient phenomenon whereas personality is more constant. In Section 7, we show for the first time that fine-grained emotion categories are useful in detecting various personality traits from stream-of-consciousness essays and collections of facebook posts.

3. CREATING EMOTION-LABELED TEXT FROM TWEETS

Sometimes people use hashtags to notify others of the emotions associated with the message they are tweeting. Table 2 shows a few examples. On reading just the message before the hashtags, most people will agree that the tweeter #1 is sad, tweeter #2 is happy, and tweeter #3 is angry.

However, there also exist tweets such as the fourth example, where reading just the message before the hashtag does not convey the emotions of the tweeter. Here, the hashtag provides information not present (implicitly or explicitly) in the rest of the message. There are also tweets, such as those shown in examples 5 and 6, that do not seem to express the emotions stated in the hashtags. This

[2]http://www.liwc.net
[3]http://people.csail.mit.edu/francois/research/personality/recognizer.html
[4]http://ota.oucs.ox.ac.uk/headers/1054.xml
[5]http://www.typealyzer.com

TABLE 2. Example tweets with emotion-words hashtags.

1. *Feeling left out... #sadness*
2. *My amazing memory saves the day again! #joy*
3. *Some jerk stole my photo on tumblr. #anger*
4. *Mika used my photo on tumblr. #anger*
5. *School is very boring today :/ #joy*
6. *to me.... YOU are ur only #fear*

may occur for many reasons including the use of sarcasm or irony. Additional context is required to understand the full emotional import of many tweets. Tweets tend to be very short, and often have spelling mistakes, short forms, and various other properties that make such text difficult to process by natural language systems. Further, it is probable, that only a small portion of emotional tweets are hashtagged with emotion words.

Our goal in this paper is to determine if we can successfully use emotion-word hashtags as emotion labels despite the many challenges outlined above:

- Can we create a large corpus of emotion-labeled hashtags?
- Are the emotion annotations consistent, despite the large number of annotators, despite no control over their socio-economic and cultural background, despite the many ways in which hashtags are used, and despite the many idiosyncracies of tweets?
- Do the hashtag annotations match with the intuitions of trained judges?

We chose to first collect tweets with hashtags corresponding to the six Ekman emotions: *#anger, #disgust, #fear, #happy, #sadness,* and *#surprise*. (Eventually, we collected tweets with hashtags corresponding to hundreds of emotions (as described in Section 6), but not before first validating the usefulness of hashtagged tweets for the basic emotions (Section 4).)

Eisenstein *et al.* (2010) collected about 380,000 tweets[6] from Twitter's official API.[7] Similarly, Go *et al.* (2009) collected 1.6 million tweets.[8] However, these datasets had less than 50 tweets that contained emotion-word hashtags. Therefore, we abandoned the search-in-corpora approach in favor of the one described below.

3.1. Hashtag-based Search on the Twitter Search API

The Archivist[9] is a free online service that helps users extract tweets using Twitter's Search API.[10] For any given query, Archivist first obtains up to 1500 tweets from the previous seven days. Subsequently, it polls the Twitter Search API every few hours to obtain newer tweets that match the query. We supplied Archivist with the six hashtag queries corresponding to the Ekman emotions, and collected about 50,000 tweets posted between November 15, 2011 and December 6, 2011.

We discarded tweets that had fewer than three valid English words. We used the *Roget Thesaurus* as the lexicon of English words.[11] This helped filter out most, if not all, of the non-English tweets that had English emotion hashtags. It also eliminated tweets that were essentially very short phrases, and tweets with very bad spelling. We discarded tweets with the prefix "Rt", "RT", and "rt", which indicate that the messages that follow are re-tweets (re-postings of tweets sent earlier by somebody else). Like González-Ibáñez *et al.* (2011), we removed tweets that did not have the hashtag of interest at the end of the message. It has been suggested that middle-of-tweet hashtags may not be good labels of the tweets.[12] Finally, we were left with about 21,000 tweets, which formed the *Hashtag Emotion Corpus (6 emotions)* or *Hashtag Emotion Corpus* for short.

[6]http://www.ark.cs.cmu.edu/GeoText
[7]https://dev.twitter.com/docs/streaming-api
[8]https://sites.google.com/site/twittersentimenthelp
[9]http://archivist.visitmix.com
[10]https://dev.twitter.com/docs/using-search
[11]Roget's Thesaurus: www.gutenberg.org/ebooks/10681
[12]End-of-message hashtags are also much more common than hashtags at other positions.

TABLE 3. Details of the Hashtag Emotion Corpus (6 emotions).

hashtag	# of instances	% of instances
#anger	1,555	7.4
#disgust	761	3.6
#fear	2,816	13.4
#joy	8,240	39.1
#sadness	3,830	18.2
#surprise	3,849	18.3
Total tweets	21,051	100.0
# of tweeters	19,059	

3.2. Distribution of emotion-word hashtags

Table 3 presents some details of the Hashtag Emotion Corpus.[13] Note that the number of tweets with emotion hashtags in the Hashtag Emotion Corpus is an order of magnitude bigger than the the number of instances in emotion-labeled headlines and blog sentences. Observe also that the distribution of emotions in the Hashtag Emotion Corpus is very different from the distribution of emotions in the 1000-Headlines Corpus (see Table 1). There are more messages tagged with the hashtag *#joy* than any of the other basic emotions.

Synonyms can often be used to express the same concept or emotion. Thus it is possible that the true distribution of hashtags corresponding to emotions is different from what is shown in Table 3. In the future, we intend to collect tweets with synonyms of *joy, sadness, fear,* etc., as well.

4. CONSISTENCY AND USEFULNESS OF HASHTAG EMOTION CORPUS

As noted earlier, even with trained judges, emotion annotation obtains only a modest inter-annotator agreement (see Table 1). As shown in Table 3, the Hashtag Emotion Corpus has about 21,000 tweets from about 19,000 different people. If the Hashtag Emotion Corpus were to be treated as manually annotated data (which in one sense, it is), then it is data created by a very large number of judges, and most judges have annotated just one instance. Therefore, an important question is to determine whether the hashtag annotations of the tens of thousands of tweeters are consistent with one another. It will also be worth determining if this large amount of emotion-tagged Twitter data can help improve emotion detection in sentences from other domains.

To answer these questions, we conducted two automatic emotion classification experiments described in the two sub-sections below. Since a particular piece of text may convey more than one emotion (it may have more than one emotion label), we use one-vs-all classifiers as they handle multi-label problems elegantly. For example, the *Anger–NotAnger* classifier may determine that the text expresses anger, and the *Disgust–NotDisgust* classifier may determine that the text conveys disgust as well. For our experiments, we created one-vs-all classifiers for each of the six basic emotions using Weka (Hall *et al.*, 2009).[14] We treated the emotion hashtags as class labels and removed them from the tweets. Thus a classifier has to determine that a tweet expresses anger, for example, without having access to the hashtag *#anger*.

We chose Support Vector Machines (SVM) with Sequential Minimal Optimization (Platt, 1999) as the machine learning algorithm because of its successful application in various research problems. We used binary features that captured the presence or absence of unigrams and bigrams. In order to set a suitable benchmark for experiments with the Hashtag Emotion Corpus, we first applied the classifiers to the SemEval-2007 Affective Text corpus. We executed ten-fold cross-validation on the 1000-Headlines dataset. We experimented with using all ngrams (one-word and two-word sequences), as well as training on only those ngrams that occurred more than once in the training data.[15]

[13]We use the number of unique user names as an approximation of the number of tweeters. It is possible that a person may have used more than one user name.

[14]http://www.cs.waikato.ac.nz/ml/weka

[15]We tokenize the text by putting white space before and after every punctuation mark.

TABLE 4. Cross-validation results on the 1000-headlines dataset. *#gold* is the number of headlines expressing a particular emotion. *#right* is the number of these instances the classifier correctly marked as expressing the emotion. *#guesses* is the number of instances marked as expressing an emotion by the classifier.

Label (X)	#gold	#right	#guesses	P	R	F
I. System using ngrams with freq. > 1						
anger	132	35	71	49.3	26.5	34.5
disgust	43	8	19	42.1	18.6	25.8
fear	247	108	170	63.5	43.7	51.8
joy	344	155	287	54.0	45.1	49.1
sadness	283	104	198	52.5	36.7	43.2
surprise	253	74	167	44.3	29.2	35.2
ALL LABELS	1302	484	912	53.1	37.2	**43.7**
II. System using all ngrams (no filtering)						
ALL LABELS	1302	371	546	67.9	28.5	40.1
III. System that guesses randomly						
ALL LABELS	1302	651	3000	21.7	50.0	30.3

The rows under I in Table 4 give a breakdown of results obtained by the *EmotionX–NotEmotionX* classifiers when they ignored single-occurrence n-grams (where X is one of the six basic emotions). *#gold* is the number of headlines expressing a particular emotion X. EmotionX is treated as the the positive class, and NotEmotionX is treated a the negative class. Therefore, *#right* is chosen to be the number of instances that the classifier correctly marked as expressing X. (Thus if the classifier correctly marks an expression with NotEmotionX, then *#right* is not affected.) *#guesses* is the number of instances marked as expressing X by the classifier. Precision (P) and recall (R) are calculated as shown below:

$$P = \frac{\#right}{\#guesses} * 100 \quad (1)$$

$$R = \frac{\#right}{\#gold} * 100 \quad (2)$$

F is the balanced F-score. The ALL LABELS row shows the sums of *#gold*, *#right*, and *#guesses*. Overall precision, recall, and F-score are calculated by plugging these values in equations 1 and 2. Thus 43.7 is the micro-average F-score obtained by these ngram classifiers.[16] The II and III rows in the table show overall results obtained by a system that uses all ngrams and by a system that guesses randomly.[17] It is not surprising that the emotion classes with the most training instances and the highest inter-annotator agreement (joy, sadness, and fear) are also the classes on which the classifiers perform best (see Table 1). We found that bigrams gave small improvements (about 1 to 2 F-score points) over and above the results obtained with unigrams alone.

The F-score of 40.1 obtained using all ngrams is close to 39.6 obtained by Chaffar and Inkpen (2011)—a sanity check for our baseline system. Ignoring words that occur only once in the training data seems beneficial. All classification results shown below are for the cases when ngrams that occurred only once were filtered out.

[16] One may choose macro-averaged F-score or micro-averaged F-score as the bottom-line statistic depending on whether one wants to give equal weight to all classes or one wants to give equal weight to each classification decision. We chose micro-average F-score since it has been a metric of choice for recent emotion analysis competitions such as the 2011 I2B2 competition on detecting emotions in suicide notes (Cherry *et al.*, 2012).

[17] A system that randomly guesses whether an instance is expressing an emotion X or not will get half of the *#gold* instances right. Further, the system will mark half of all the instances as expressing emotion X. For ALL LABELS,
$\#right = \frac{\#gold}{2}$, and $\#guesses = \frac{\#instances*6}{2}$.

TABLE 5. Cross-validation results on the Hashtag Emotion Corpus. The highest F-score is shown in bold.

Label	#gold	#right	#guesses	P	R	F
I. System using ngrams with freq. > 1						
anger	1555	347	931	37.3	22.31	27.9
disgust	761	102	332	30.7	13.4	18.7
fear	2816	1236	2073	59.6	43.9	50.6
joy	8240	4980	7715	64.5	60.4	62.4
sadness	3830	1377	3286	41.9	36.0	38.7
surprise	3849	1559	3083	50.6	40.5	45.0
ALL LABELS	21051	9601	17420	55.1	45.6	**49.9**
II. System that guesses randomly						
ALL LABELS	21051	10525	63,153	16.7	50.0	21.7

TABLE 6. Results on the 250-headlines dataset. Highest F-scores in I and II are shown in bold.

	# of features	P	R	F
I. System using ngrams in training data:				
a. the 1000-headlines text (target domain)	1,181	40.2	32.1	35.7
b. the Hashtag Emotion Corpus (source domain)	32,954	29.9	26.1	27.9
c. the 1000-headlines text and the Hashtag Emotion Corpus (target and source)				
c.1. no domain adaptation	33,902	41.7	35.5	38.3
c.2. with domain adaptation	101,706	46.0	35.5	**40.1**
II. System using ngrams in 1000-headlines and:				
a. the Hashtag Emotion Lexicon	1,181 + 6	44.4	35.3	39.3
b. the WordNet Affect Lexicon	1,181 + 6	39.7	30.5	34.5
c. the NRC Emotion Lexicon	1,181 + 10	46.7	38.6	**42.2**
III. System that guesses randomly	-	27.8	50.0	35.7

4.1. Experiment I: Can a classifier learn to predict emotion hashtags?

We applied the binary classifiers described above to the Hashtag Emotion Corpus. Table 5 shows the ten-fold cross-validation results. Observe that even though the Hashtag Emotion Corpus was created from tens of thousands of users, the automatic classifiers are able to predict the emotions (hashtags) with F-scores much higher than the random baseline, and also higher than those obtained on the 1000-headlines corpus. Note also that this is despite the fact that the random baseline for the 1000-headlines corpus ($F = 30.3$) is higher than the random baseline for the Hashtag Emotion Corpus ($F = 21.7$). The results suggest that emotion hashtags assigned to tweets are consistent to a degree such that they can be used for detecting emotion hashtags in other tweets.

Expectedly, the *Joy–NotJoy* classifier gets the best results as it has the highest number of training instances. The *Sadness–NotSadness* classifier performed relatively poorly considering the amount of training instances available, whereas the *Fear-NotFear* classifier performed relatively well. It is possible that people use less overt cues in tweets when they are explicitly giving it a sadness hashtag.

4.2. Experiment II: Can the Hashtag Emotion Corpus help improve emotion classification in a different domain?

Usually, supervised algorithms perform well when training and test data are from the same domain. However, domain adaptation algorithms may be used to combine training data in the target domain with large amounts of training data from a different source domain. Several successful domain adaptation techniques have been proposed in the last few years including those by Daumé (2007), Blitzer *et al.* (2007), and Ganchev *et al.* (2012). We used the one proposed by Daumé (2007) in our experiments, mainly because of its simplicity.

The Daumé (2007) approach involves the transformation of the original training instance feature vector into a new space made up of three copies of the original vector. The three copies correspond to the target domain, the source domain, and the general domain. If X represents an original feature

vector from the *target domain*, then it is transformed into XOX, where O is a zero vector. If X represents original feature vector from the *source domain*, then it is transformed into OXX. This data is given to the learning algorithm, which learns information specific to the target domain, specific to the source domain, as well as information that applies to both domains. The test instance feature vector (which is from the target domain) is transformed to XOX. Therefore, the classifier applies information specific to the target domain as well as information common to both the target and source domains, but not information specific only to the source domain.

In this section, we describe experiments on using the Hashtag Emotion Corpus for emotion classification in the newspaper headlines domain. We applied our binary emotion classifiers on unseen test data from the newspaper headlines domain—the 250-Headlines dataset—using each of the following as a training corpus:

- Target-domain data: the 1000-Headlines data.
- Source-domain data: the Hashtag Emotion Corpus.
- Target and Source data: A joint corpus of the 1000-Headlines set and the Hashtag Emotion Corpus.

Additionally, when using the 'Target and Source' data, we also tested the domain adaptation algorithm proposed in Daumé (2007). Since the *EmotionX* class (the positive class) has markedly fewer instances than the *NotEmotionX* class, we assigned higher weight to instances of the positive class during training.[18] The rows under I in Table 6 give the results. (Row II results are for the experiment described in Section 6, and can be ignored for now.)

The micro-averaged F-score when using target-domain data (row I.a.) is identical to the score obtained by the random baseline (row III). However, observe that the precision of the ngram system is higher than the random system, and its recall is lower. This suggests that the test data has many n-grams not previously seen in the training data. Observe that as expected, using source-domain data produces much lower scores (row I.b.) than when using target-domain training data (row I.a.).

Using both target- and source-domain data produced significantly better results (row I.c.1.) than using target-domain data alone (I.a.). Applying the domain adaptation technique described in Daumé (2007), obtained even better results (row I.c.2.). The use of Hashtag Emotion Corpus improved both precision and recall over just using the target-domain text. This shows that the Hashtag Emotion Corpus can be leveraged, preferably with a suitable domain adaptation algorithm, to improve emotion classification results even on datasets from a different domain. It is also a validation of the premise that the self-labeled emotion hashtags are consistent, at least to some degree, with the emotion labels given by trained human judges.

5. CREATING THE HASHTAG EMOTION LEXICON

Word–emotion association lexicons are lists of words and associated emotions. For example, the word *victory* may be associated with the emotions of joy and relief. These emotion lexicons have many applications, including automatically highlighting words and phrases to quickly convey regions of affect in a piece of text. Mohammad (2012b) shows that these lexicon features can significantly improve classifier performance over and above that obtained using ngrams alone.

WordNet Affect (Strapparava and Valitutti, 2004) includes 1536 words with associations to the six Ekman emotions.[19] Mohammad and colleagues compiled emotion annotations for about 14,000 words by crowdsourcing to Mechanical Turk (Mohammad and Turney, 2013; Mohammad and Yang, 2011).[20] This lexicon, referred to as the NRC Emotion Lexicon, has annotations for eight emotions (six of Ekman, trust, and anticipation) as well as for positive and negative sentiment.[21] Here, we show how we created an ngram–emotion association lexicons from emotion-labeled sentences in the Hashtag Emotion Corpus.

[18] For example, for the *anger–NotAnger* classifier, if 10 out of 110 instances have the label anger, then they are each given a weight of 10, whereas the rest are given a weight of 1.

[19] http://wndomains.fbk.eu/wnaffect.html

[20] http://www.purl.org/net/saif.mohammad/research

[21] Plutchik (1985) proposed a model of 8 basic emotions.

TABLE 7. Number of word types in emotion lexicons.

Emotion lexicon	# of word types
1000-Headlines Lexicon (6 emotions)	152
Hashtag Emotion Lexicon (6 emotions)	11,418
WordNet Affect Lexicon (6 emotions)	1,536
NRC Emotion Lexicon (8 emotions)	14,000

5.1. Method

Given a dataset of sentences and associated emotion labels, we compute the *Strength of Association (SoA)* between an n-gram w and an emotion e to be:

$$SoA(w, e) = PMI(w, e) - PMI(w, \neg e) \tag{3}$$

where PMI is the pointwise mutual information.

$$PMI(w, e) = \log_2 \frac{freq(w, e) * N}{freq(w) * freq(e)} \tag{4}$$

where $freq(w, e)$ is the number of times w occurs in a sentence with label e.
$freq(w)$ and $freq(e)$ are the frequencies of w and e in the labeled corpus.
N is the number of words in the dataset.

$$PMI(w, \neg e) = \log_2 \frac{freq(w, \neg e) * N}{freq(w) * freq(\neg e)} \tag{5}$$

where $freq(w, \neg e)$ is the number of times n occurs in a sentence that does not have the label e. $freq(\neg e)$ is the number of sentences that do not have the label e. Thus, equation 4 is simplified to:

$$SoA(w, e) = \log_2 \frac{freq(w, e) * freq(\neg e)}{freq(e) * freq(w, \neg e)} \tag{6}$$

Since PMI is known to be a poor estimator of association for low-frequency events, we ignore ngrams that occur less than five times.

If an n-gram has a stronger tendency to occur in a sentence with a particular emotion label, than in a sentence that does not have that label, then that ngram–emotion pair will have an SoA score that is greater than zero.

5.2. Emotion lexicons created from the 1000-headlines dataset and the Hashtag Emotion Corpus

We calculated SoA scores for the unigrams and bigrams in the Hashtag Emotion Corpus with the six basic emotions. All ngram–emotion pairs that obtained scores greater than zero were extracted to from the *Hashtag Emotion Lexicon (6 emotions)* (*Hashtag Lexicon* for short). We repeated these steps for the 1000-headlines dataset as well. Table 7 shows the number of word types in the two automatically generated and the two manually created lexicons (WordNet Affect and NRC Emotion lexicon). Observe that the 1000-headlines dataset produces very few entries, whereas the large size of the Hashtag Emotion Corpus enables the creation of a substantial emotion lexicon.

The Hashtag Emotion Lexicon should not be confused with the NRC Emotion Lexicon. The NRC Emotion Lexicon was created in 2011 by manual annotation through Mechanical Turk, whereas the Hashtag Lexicon was created automatically from tweets with emotion-word hashtags posted in 2012. The NRC Emotion Lexicon has binary association score (0 or 1), whereas the Hashtag Lexicon provides real-valued scores that indicate the degree of word–emotion association (higher scores imply higher association).

5.3. Evaluating the Hashtag Emotion Lexicon

We evaluate the Hashtag Emotion Lexicon by using it for classifying emotions in a setting similar to the one discussed in the previous section. The test set is the 250-headlines dataset. The training set is the 1000-headlines dataset. We used binary features that captured the presence or absence of

unigrams and bigrams just as before. Additionally, we also used integer-valued affect features that captured the number of word tokens in a sentence associated with different emotions labels in the Hashtag Emotion Lexicon and the WordNet Affect Lexicon. For example, if a sentence has two joy words and one surprise word, then the joy feature has value 2, surprise has value 1, and all remaining affect features have value 0.[22]

We know from the results in Table 6 (I.a. and I.c) that using the Hashtag Emotion Corpus in addition to the 1000-headlines training data significantly improves results. Now we investigate whether the Hashtag Emotion Lexicon can similarly improve performance. The rows under II in Table 6 give the results.

Observe that even though the Hashtag Emotion Lexicon is a derivative of the hashtag Emotion Corpus that includes fewer unigrams and bigrams, the classifiers using the Hashtag Emotion Lexicon produces an F-score (row II.a.) significantly higher than in the scenarios of I.a. and almost as high as in I.c.2. This shows that the Hashtag Emotion Lexicon successfully captures the word–emotion associations that are latent in the Hashtag Emotion Corpus. We also find that the classifiers perform significantly better when using the Hashtag Emotion Lexicon (row II.a.) than when using the WordNet Affect Lexicon (row II.b.), but not as well as when using the NRC Emotion Lexicon (row II.c.). The strong results of the NRC Emotion Lexicon are probably because of its size and because it was created by manual annotation of words for emotions, which required significant time and effort. On the other hand, the Hashtag Emotion Lexicon can be easily improved further by compiling an even larger set of tweets using synonyms and morphological variants of the emotion words used thus far.

6. EXTENDING THE EMOTION-LABELED DATASET AND THE EMOTION LEXICON TO INCLUDE HUNDREDS OF FINE EMOTION CATEGORIES

The previous sections show that emotion-word hashtagged tweets are a good source of labeled data for automatic emotion processing. Those experiments were conducted using tweets pertaining to the six Ekman emotions because labeled evaluation data exists for only those emotions. However, a significant advantage of using hashtagged tweets is that we can collect large amounts of labeled data for any emotion that is used as a hashtag by tweeters. Thus we polled the Twitter API and collected a large corpus of tweets pertaining to a few hundred emotions.

We used a list of 585 emotion words compiled by Zeno G. Swijtink as the hashtagged query words.[23] Note that we chose not to dwell on the question of whether each of the words in this set is truly an emotion or not. Our goal was to create and distribute a large set of affect-labeled data, and users are free to choose a subset of the data that is relevant to their application. The final corpus, which we call the *Hashtag Emotion Corpus (585 emotions)*, has 2,914,085 tweets in all, each with at least one of the 585 emotion words as a hashtagged word. This is the first text collection labeled for hundreds of emotions.

We generated a word–emotion association lexicon from these 2,914,085 tweets just as described earlier in Section 5. We call it the *Hashtag Emotion Lexicon (585 emotions)*. In total, it has 15,825 entries, where as entry is a word–emotion pair and an association score. This is the first word-emotion association lexicon that has real-valued association scores for hundreds of emotions (previous lexicons involved only a handful of emotions, and often had 0 or 1 association values).

For the rest of the paper, we will only make use of the Hashtag Emotion Corpus (585 emotions) and the Hashtag Emotion Lexicon (585 emotions), and not the 6-emotions version described in Sections 3, 4, and 5. Thus, we will refer to the Hashtag Emotion Corpus (585 emotions) simply as the Hashtag Emotion Corpus and the Hashtag Emotion Lexicon (585 emotions) simply as the Hashtag Emotion Lexicon for short. We will now show the use of the Hashtag Emotion Lexicon in an extrinsic task—detecting personality from written text.

[22]Normalizing by sentence length did not give better results.

[23]http://www.sonoma.edu/users/s/swijtink/teaching/philosophy_101/paper1/listemotions.htm

7. DETECTING PERSONALITY USING FINE EMOTION CATEGORIES

We investigate the relationship between emotions and personality using the Hashtag Emotion Lexicon (with its hundreds of emotion associations) over a personality detection task. The goal of personality detection from text is to automatically analyze free-form text written by a person in order to infer her personality traits, such as extroversion and neuroticism.

We detect personality in two datasets: stream-of-consciousness essays and collections of Facebook status updates. The Essays dataset was collected by Pennebaker and King (1999). It consists of 2,469 essays (1.9 million words) by psychology students. The Facebook dataset is a collection of 10,000 Facebook posts (status updates) by 250 users. We concatenated all status updates by a user into a single text that was fed to our system. Both datasets were provided as part of a shared task in the Workshop on Computational Personality Detection.[24] Personality was assessed by asking the students to respond to a Big Five Inventory Questionnaire (John and Srivastava, 1999).

We first build a number of baseline classifiers that rely on commonly used features for personality detection. Some of these baselines use coarse affect features about evaluativeness (positive and negative sentiment) and coarse emotion categories (the six basic emotion categories). We then add features drawn from the Hashtag Emotion Lexicon to determine if using features pertaining to hundreds of fine emotions improves performance over an above the baselines. The subsections below describe some of the lexicons used to obtain features for baseline systems, the personality classification system (classifier and features used), and our experiments on the two datasets.

7.1. Lexicons used to obtain features for the baseline systems

7.1.1. *Specificity Lexicon.* Gill and Oberlander (2002), and later Mairesse *et al.* (2007), show that people with a neurotic personality tend to use concrete words more frequently. Inspired by this, we explore if people of a certain personality type tend to use terms with high specificity. The specificity of a term is a measure of how general or specific the referred concept is. For example, *entity* is a very general concept whereas *ball-point pen* is a very specific concept.

Resnik (1995) showed that specificity or information content of WordNet synsets can be accurately determined by using corpus counts. Pedersen pre-computed information content scores for 82,115 WordNet noun synsets and 13,708 verb synsets using the British National Corpus (BNC). We created a word-level information content lexicon by first mapping the words to their synsets, and then assigning the words with information content scores of the corresponding synsets. If a word is associated with more than one synset, then the synset with the highest information content is chosen. The final lexicon had 66,464 noun entries and 6,439 verb entries. In contrast with the Hashtag Lexicon, which has fine emotion categories, the specificity lexicon captures how general or specific a word is, without regard to whether it is associated with an emotion or not. Thus a comparative experiment with the specificity features sheds light on whether the Hashtag Lexicon is useful because of information from fine *emotion* categories or simply because of information from fine categories (emotional or otherwise). We computed the average information content of the words in the input text and used it as a feature in our machine learning system.

7.1.2. *Coarse Affect Lexicon 1: Osgood Lexicon.* Osgood *et al.* (1957) asked human subjects to rate words on various scales such as complete–incomplete, harmonious–dissonant, and high–low. They then performed a factor analysis of these ratings to discover that most of the variation was due to three dimensions: evaluativeness (*good–bad*), activity (*active–passive, large–small*), and potency (*sharp–dull, fast–slow*). Turney and Littman (2003) proposed a method to automatically calculate a word's evaluativeness score using a vector space model and word–word co-occurrence counts in text. Turney later generated lexicons of word–evaluativeness scores and additionally lexicons of word–activity and word–potency scores for 114,271 words from WordNet. In contrast with the Hashtag Lexicon, which has hundreds of fine-grained affect categories, the Osgood Lexicon has only three coarse affect categories (evaluativeness, activity, and potency). We used these lexicons and computed the average evaluativeness, activity, and potency scores of the words in the text to be classified.

[24] http://mypersonality.org/wiki/doku.php?id=wcpr13

7.1.3. *Coarse Affect Lexicon 2: NRC Emotion Lexicon.* The NRC Emotion Lexicon has about 14,000 words annotated for eight emotions (six of Ekman, trust, and anticipation) as well as for positive and negative sentiment. In contrast with the Hashtag Lexicon, which has hundreds of fine-grained affect categories, the NRC Emotion Lexicon has eight coarse categories. For each of the eight emotions, we used as features the average number of emotion words in the text to be classified.

7.1.4. *Emotion Clusters Lexicon.* Consider a multi-dimensional space where each of the words in the vocabulary is a dimension. A fine-grained emotion category can be positioned in this space using the list of words and real-valued association scores with the emotion (as in the Hashtag Lexicon). We calculated the similarity between two emotions in this space by the cosine of the two corresponding vectors. Finally, we clustered the 585 emotions into eight groups by repeated bisection using the clustering tool CLUTO (Karypis, 2003). We then created a word–cluster association lexicon from tweets where the hashtagged fine emotion category was replaced by the corresponding cluster (just as described in Section 5). In contrast with the Hashtag Lexicon and just as the NRC Emotion Lexicon, the word–cluster lexicon has eight coarse categories. However, unlike the NRC Emotion Lexicon, the word–cluster lexicon has the same vocabulary as the Hashtag Lexicon. Thus a comparative experiment with the cluster lexicon tells us more clearly whether fine-emotion categories are useful for personality detection. We used as features, the average of the cluster association score for each of the eight emotion clusters in the input text.

7.2. System Description

Classifier: We trained five binary Support Vector Machine (SVM) classifiers, one for each of the five personality dimensions: extroversion (EXT), neurotism (NEU), agreeability (AGR), conscientiousness (CON), and openness (OPN). SVM is a state-of-the-art learning algorithm proven to be effective on text categorization tasks and robust on large feature spaces. In each experiment, the results were averaged over ten runs of stratified cross-validation. We used the LibSVM package (Chang and Lin, 2011) with a linear kernel.

The essays dataset is well-balanced in the number of instances in each class for each of the personality dimensions. (The highest skew is 53%–47% for the AGR–NotAGR dimension.) However, the Facebook posts dataset is much more skewed. (For example, 70%–30% for the OPN–NotOPN dimension.) Thus we used the essays dataset as is, but employed the following strategy for the Facebook posts. During each of the ten cross-validation runs, for each personality trait, we randomly select N instances of a majority class, where N is the number of instances in the minority class.

Features: Each input text was represented by the following groups of features:

a. Mairesse Baseline (MB): This is the complete set of features used by Mairesse *et al.* (2007). Some of these features are listed below: word count, words per sentence, type/token ratio, words longer than six letters, negations, assents, articles, prepositions, numbers, pronouns (first person, second person, third person), emotion words, cognition words (*insight, tentative*), sensory and perceptual words (*see, hear*), social processes words (*chat, friend*), time words, space words, motion words, punctuations, and swear words.
b. Token unigrams: Frequencies of tokens divided by the total number of tokens in an input text.
c. Average Information Content: Average information content of the text, calculated using the Specificity Lexicon.
d. Features from the Osgood Lexicon - CoarseAff (Osgood Lexicon): Average potency of the text, average evaluativeness of the text, and the average activity score of the text.
e. Features from the NRC Emotion Lexicon - CoarseAff (NRC Emotion Lexicon): Average number of emotion words for each of the 8 emotions in the NRC Emotion Lexicon.
f. Features from Hashtag Emotion Lexicon - FineEmo (Hashtag Emotion Lexicon): Average of the emotion association score for each of the 585 emotions in the Hashtag Lexicon.
g. Features from the Cluster Lexicon - CoarseAff (Cluster Lexicon): Average of the cluster association score for each of the 8 clusters in the input text.

Table 8. Accuracy of automatic classification of essays into the big five dimensions of personality. *Combination* is the system that combines exactly those features that provide significant improvement over the majority classifier baseline when used individually. The features combined for the five dimensions are as follows: for EXT: a + f; for NEU: a + d + e + f; for AGR: a + f; for CON: a + d + e + f; and for OPN: a + d + f. All statistically significant improvements ($p < .05$) over the majority baseline are marked with a *. All statistically significant improvements ($p < .05$) over the Mairesse Baseline are shown in bold.

	EXT	NEU	AGR	CON	OPN
Majority Classifier	51.74	50.04	53.08	50.81	51.54
SVM Classifier					
a. Mairesse Baseline (MB)	55.13*	58.09*	55.35*	55.28*	59.57*
b. Unigrams	51.74	50.04	53.08	50.81	51.54
c. Average Information Content	51.74	50.09	53.08	50.64	51.54
d. CoarseAff (Osgood Lexicon)					
Activity	51.68	53.76*	53.04	51.78*	54.61*
Evaluative	51.67	52.78*	53.01	54.81*	52.20*
Potency	51.77	53.31*	53.06	51.62*	52.61*
All three	51.58	54.52*	52.95	54.38*	57.32*
e. CoarseAff (NRC Emotion Lexicon)	51.74	51.05*	53.10*	51.28*	51.54
f. FineEmo (Hashtag Emotion Lexicon)	55.29*	55.75*	56.03*	**56.54***	**60.68***
g. MB + CoarseAff					
Activity	55.25*	58.21*	55.33*	55.20*	59.49*
Evaluative	55.39*	58.12*	55.32*	55.44*	59.42*
Potency	55.25*	**58.33***	55.25*	55.27*	59.64*
All three	55.34*	**58.31***	55.15*	55.51*	59.64*
h. MB + CoarseAff (NRC Emotion Lexicon)	55.24*	**58.33***	55.37*	55.45*	59.60*
i. MB + FineEmo (Hashtag Emotion Lexicon)	**56.45***	58.26*	55.13*	**56.73***	**60.64***
j. Combination	**56.45***	58.04*	55.13*	**56.68***	**60.62***
k. FineEmo-clusters	51.74	50.97*	53.09	52.11*	58.11*

7.3. Experiments on the Essays Dataset

7.3.1. *Results.* Upon classification, the results were compared with the gold labels of yes or no for each of the five personality dimensions. Table 8 shows the accuracies of the yes and no labels for the five personality classes extroversion (EXT), neurotism (NEU), agreeability (AGR), conscientiousness (CON), and openness (OPN). We also present the results for a simple baseline classifier that always predicts the majority class. All statistically significant improvements ($p < 0.05$) over the majority baseline are marked with a *. All statistically significant improvements ($p < 0.05$) over the Mairesse Baseline are shown in bold.

The Mairesse Baseline performs significantly better than the majority baseline for all five personality dimensions. Since the Mairesse Baseline is one of the well-known baselines in personality detection, we compare performance of other feature groups with this baseline and show all results that are significantly higher than it in bold. Unigrams and average information content features fail to improve results over the majority baseline. All three of the Osgood Lexicon features improve results significantly over the majority baseline for the NEU, CON, and OPN classes (d rows), however, when combined with the Mairesse et al. features, only the potency features improve results significantly over the Mairesse Baseline, and that too only for NEU (g rows). The NRC Emotion lexicon features lead to statistically significant improvements over the majority baseline for the NEU, AGR and CON classes (row e), however, when added to the Mairesse et al. features, they improve results significantly over the Mairesse Baseline again only for NEU (row h). In contrast to the coarse affect features of Osgood and NRC Emotion lexicon, the fine emotion features of the Hashtag Emotion Lexicon lead to significant improvements over the majority baseline in all five personality classes. Observe also

that the Hashtag Lexicon features alone obtain higher results than the collection of features used in the Mairesse Baseline for all personality dimensions, except NEU. Adding these features on top of the Mairesse Baseline leads to significant improvements in the EXT, CON, and OPN classes. These results show that the fine emotion categories from the Hashtag Lexicon are a particularly useful source of features for detecting personality from written text.

Combination (row j) is the system that combines exactly those features that provide significant improvement over the majority classifier baseline when used individually. The features combined for the five dimensions are as follows: for EXT: rows a + f; for NEU: rows a + d + e + f; for AGR: rows a + f; for CON: rows a + d + e + f; and for OPN: rows a + d + f. Observe that the combinations do not lead to better results than those obtained using just the Mairesse Baseline and the Hashtag Lexicon.

In order to determine whether the benefit from the Hashtag Lexicon is due to the grouping of terms into fine categories or simply because of its vocabulary coverage, we present the results obtained using the lexicon created by first clustering the emotion hashtags into eight coarse categories (row k). We find that the results obtained using the Hashtag Lexicon (row f) are significantly better than those obtained using the cluster lexicon (row k) for all five personality dimensions.

7.3.2. *Discussion.* The fact that unigram features are not as helpful as in some other tasks such as classification of text by topic, is one of the reasons personality detection is a relatively hard problem. Nonetheless, the fine-grained emotion features from the Hashtag Lexicon provided statistically significant gain over the baseline. In contrast, coarse affect features and specificity features failed to provide the same amount of improvements. This suggests that fine affect categories contain useful discriminating information not present in coarse affect categories or simple specificity features.

In order to identify which of the 585 emotions had the most discriminative information, we calculated information gain of each of 585 emotion features. (Decision tree learners use information gain to determine the sequence of nodes in the tree.) Table 9 shows the top ten emotion categories with the highest gain for the five personality dimensions. Observe that most of the emotions seem to be reasonable indicators of the corresponding personality trait. Note that the columns include emotions that are indicative of either of the two ends of the personality dimensions (for example, the emotions in column EXT are associated with either extroversion or introversion). Observe also that some of these emotions are very close to the basic emotions of happiness and sadness, but many are emotions felt at relatively specific situations, such as guilt, excitement, anxiety, and shame.

The emotion categories at the bottom of the information gain lists (not shown here) are either (1) not relevant for discriminating the target personality dimension, and/or (2) lacking sufficient information in the Hashtag Lexicon. For example, we found very few tweets with #petulant, #belittled, and #genial.

The five terms most associated with the lexical categories of *#possessive* and *#apart* (the two most discriminative emotion categories for EXT) are shown below:

#possessive: *possessive*: 7.228, *hottie*: 6.448, *tense*: 5.911, *lover*: 5.213, *mine*: 4.141, ...

#apart: *apart*: 4.6, *tear*: 4.065, *miss*: 2.341, *fall*: 2.085, *heart*: 1.63, ...

The numbers next to the words are their PMI scores with the emotion word hashtag. Observe that the terms in the #possessive category tend to be used more often by an extrovert, whereas the terms in the #apart category tend to be associated more with introverts.

7.4. Experiments on the Facebook Dataset

7.4.1. *Results.* Table 10 shows the accuracies obtained on the Facebook dataset. All statistically significant improvements ($p < 0.05$) over the majority baseline are marked with a *. All statistically significant improvements ($p < 0.05$) over the Mairesse Baseline are shown in bold. The baseline classifier that always predicts the majority class gets 50% accuracy on this dataset. Observe that here, the Mairesse Baseline provided statistically significant improvements over the majority baseline only for the EXT, CON, and OPN classes. Once again, unigrams are not useful. The average information

Table 9. Essays Dataset: Top ten hashtag emotion categories with highest information gain for personality classification.

EXT	NEU	AGR	CON	OPN
#possessive	#guilt	#happy	#excited	#anxious
#apart	#eager	#anger	#apprehensive	#delighted
#happy	#interested	#homesick	#anger	#blah
#cherish	#keen	#giddy	#hate	#exhausted
#admiring	#helpless	#chaotic	#ashamed	#sweet
#impaired	#passion	#heartbroken	#giddy	#tired
#jealousy	#unhappy	#sweet	#partial	#lonely
#gleeful	#insignificant	#neglected	#disturbed	#nervous
#vibrant	#timid	#loving	#wrecked	#ecstatic
#huggy	#anticipation	#lonely	#needed	#wrecked

Table 10. Accuracy of automatic classification of Facebook status updates into the big five dimensions of personality. *Combination* is the system that combines exactly those features that provide significant improvement over the majority classifier baseline when used individually. The features combined for the five dimensions are as follows: for EXT: a + d + e + f; for NEU: c + e + f; for AGR: d + f; for CON: a + f; and for OPN: a + d + e + f. All statistically significant improvements ($p < .05$) over the Majority Baseline are marked with a *. All statistically significant improvements ($p < .05$) over Mairesse Baseline are shown in bold.

	EXT	NEU	AGR	CON	OPN
Majority Classifier	50.00	50.00	50.00	50.00	50.00
SVM Classifier					
a. Mairesse Baseline (MB)	54.16*	48.72	51.04	53.12*	53.65*
b. Unigrams	49.26	49.90	50.35	49.46	49.59
c. Average Information Content	50.05	**54.85***	49.61	50.38	50.27
d. CoarseAff (Osgood Lexicon)	53.16*	49.95	**54.35***	49.75	52.97*
e. BasicEmo (NRC Emotion Lexicon)	52.16*	**54.34***	50.52	49.92	54.39*
f. FineEmo (Hashtag Emotion Lexicon)	52.00*	**53.21***	**59.22***	52.12*	53.72*
g. MB + FineEmo	54.53*	48.32	49.22	50.50	53.51*
h. Combination	55.37*	**56.63***	**58.91***	50.50	53.92*
i. FineEmo-clusters	49.68	**51.79**	51.96*	**59.46***	53.38*

content helps only NEU, whereas the Osgood features improve classification of EXT, AGR, and OPN, and the NRC Emotion Lexicon features improve classification of EXT, NEU, and OPN.

Once again, in contrast to the coarse affect features of Osgood and NRC Emotion lexicon, the fine emotion features of the Hashtag Emotion Lexicon lead to significant improvements over the majority baseline in all five personality classes. The improvement for AGR is especially large. Observe also that the Hashtag Lexicon features alone obtain higher results than the collection of features used in the Mairesse Baseline for all personality dimensions, except EXT and CON (the improvements are statistically significant for NEU and AGR). Using both the Mairesse Baseline and the Hashtag lexicon (row g), however, does not lead to better results. *Combination* (row h) is the system that combines exactly those features that provide significant improvement over the majority classifier baseline when used individually. It beats the Mairesse Baseline on NEU and AGR. Results obtained using the Hashtag Lexicon (row f) are better than those obtained using the cluster lexicon (row i) for EXT, NEU, AGR, and OPN, but interestingly, for CON, using the cluster lexicon is markedly helpful.

TABLE 11. Facebook Dataset: Top ten hashtag emotion categories with highest information gain for personality classification.

EXT	NEU	AGR	CON	OPN
#unimportant	#happiness	#like	#calm	#tranquil
#attached	#bugged	#crushed	#considerate	#lust
#destroyed	#alert	#mixed	#mean	#love
#detached	#irate	#delighted	#cowardly	#appreciative
#awful	#cold	#happiness	#bewildered	#peaceful
#lust	#anxious	#hopeful	#shunned	#jealousy
#comfortable	#weak	#blushing	#exposed	#innerpeace
#burned	#shame	#jealousy	#imperfect	#thoughtful
#troubled	#sexy	#grateful	#judged	#touched
#jumpy	#desire	#spirited	#pity	#careful

7.4.2. *Discussion.* In order to identify which of the 585 emotions had the most discriminative information on the Facebook dataset, we again calculated information gain of each of 585 emotion features. Table 11 shows the top ten emotion categories with the highest gain for the five personality dimensions. Since the dimension of agreeability–disagreeability (AGR) was most helped by the hashtag lexicon, we list below some of the terms most associated with the lexical categories of *#like* and *#crushed* (the two most discriminative emotion categories for AGR):

#like: *hash*: 3.72, *likes*: 2.361, *word*: 1.682, *picture*: 1.639, *wit*: 1.58, *ice*: 1.38, *hey*: 1.379, *twitter*: 1.214, *tweet*: 1.176, *cool*: 1.097, *follow*: 1.057, *nice*: 0.934, *song*: 0.887, *people*: 0.873, *fun*: 0.831, ...

#crushed: *crushed*: 3.444, *estimated*: 3.436, *heartbroken*: 3.186, *devastating*: 2.968, *hurtful*: 2.564, *heartbreaking*: 2.353, *santa*: 2.312, *childhood*: 2.295, *disappointed*: 2.076, *upsetting*: 2.047, *goodbye*: 2.023, ...

The numbers next to the words are their PMI scores with the emotion word hashtag. Observe that the terms in the #like category tend to be used more often by an agreeable person, whereas the terms in the #crushed category tend to be associated more with those that not as much.

8. CONCLUSIONS AND FUTURE WORK

We compiled a large corpus of tweets, the Hashtag Emotion Corpus, labeled with hundreds of fine emotion categories using emotion-word hashtags. Even though the corpus has tweets from several thousand people, we showed that the self-labeled hashtag annotations are consistent. We also showed how the Hashtag Emotion Corpus can be combined with labeled data from a different target domain to improve automatic classification accuracy. This experiment was especially telling since it showed that self-labeled emotion hashtags correspond well with annotations of trained human judges.

We extracted a large word–emotion association lexicon, the Hashtag Emotion Lexicon, from the Hashtag Emotion Corpus. This is the first lexicon with association information for hundreds of emotions. It also has real-valued association score that indicate the degree of association. We showed that the Hashtag Emotion Lexicon is of good quality by using the sentence classification task as a test bed, where classifiers using it performed significantly better than those that used the manually created WordNet Affect lexicon.

We performed experiments on personality detection from text using the Hashtag Emotion Lexicon to established a relation between emotions and personality. Specifically, we showed that lexical categories corresponding to fine-grained emotions such as excitement, guilt, yearning, and admiration (extracted from the Hashtag Lexicon) are valuable features in the detection of personality. We performed experiments using three large automatically created lexicons of fine emotion categories, coarse affect categories, and word information content. The fine emotion category features extracted form the Hashtag Emotion Lexicon significantly improved performance of all classifiers over the majority baseline, and match or outperform a known set of baseline features—Mairesse *et al.* (2007).

The improvements were in large majority of cases above and beyond those obtained using features from coarse affect categories and word information content.

Our future work includes collecting emotion-word hashtagged tweets in other languages such as Spanish and Arabic. We also want to collect tweets with hashtags that are near-synonyms of the emotion terms described in this paper. We want to determine if there is a difference in emotions associated with different morphological forms of the emotion words, for example, *#sad* versus *#sadness* or *#anxious* versus *#anxiety*. There are several applications that can benefit from the data collected in this paper including early depression detection, therapeutic benefits of expressing emotions, tracking public sentiment towards commercial products, and identifying how people use emotional expression in microblogs to persuade others. All resources created by the authors and used in this research effort, including the Hashtag Lexicon, are freely available.[25]

REFERENCES

Alm, C. O. and Sproat, R. (2005). Emotional sequencing and development in fairy tales. In *Proceedings of the First International Conference on Affective Computing and Intelligent Interaction*, pages 668–674.

Alm, C. O., Roth, D., and Sproat, R. (2005). Emotions from text: Machine learning for text-based emotion prediction. In *Proceedings of the Joint Conference on HLT–EMNLP*, Vancouver, Canada.

Aman, S. and Szpakowicz, S. (2007). Identifying expressions of emotion in text. In V. Matoušek and P. Mautner, editors, *Text, Speech and Dialogue*, volume 4629 of *Lecture Notes in Computer Science*, pages 196–205. Springer Berlin / Heidelberg.

Blitzer, J., Dredze, M., and Pereira, F. (2007). Biographies, bollywood, boomboxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*, pages 187–205, Prague, Czech Republic.

Bollen, J., Mao, H., and Pepe, A. (2011). Modeling Public Mood and Emotion: Twitter Sentiment and Socio-Economic Phenomena. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, pages 450–453, Barcelona, Spain.

Boucouvalas, A. C. (2002). Real time text-to-emotion engine for expressive internet communication. *Emerging Communication: Studies on New Technologies and Practices in Communication*, **5**, 305–318.

Bougie, J. R. G., Pieters, R., and Zeelenberg, M. (2003). Angry customers don't come back, they get back: The experience and behavioral implications of anger and dissatisfaction in services. Open access publications from tilburg university, Tilburg University.

Chaffar, S. and Inkpen, D. (2011). Using a heterogeneous dataset for emotion analysis in text. In *Canadian Conference on AI*, pages 62–67.

Chang, C.-C. and Lin, C.-J. (2011). LIBSVM: A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology*, **2**, 27:1–27:27.

Cherry, C., Mohammad, S. M., and de Bruijn, B. (2012). Binary classifiers and latent sequence models for emotion detection in suicide notes. *Biomedical Informatics Insights*, **5**, 147–154.

Daumé, H. (2007). Frustratingly Easy Domain Adaptation. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*, Prague, Czech Republic.

Eisenstein, J., O'Connor, B., Smith, N. A., and Xing, E. P. (2010). A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1277–1287, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion*, **6**(3), 169–200.

Francisco, V. and Gervás, P. (2006). Automated mark up of affective information in english texts. In P. Sojka, I. Kopecek, and K. Pala, editors, *Text, Speech and Dialogue*, volume 4188 of *Lecture Notes in Computer Science*, pages 375–382. Springer Berlin / Heidelberg.

[25]Email Saif Mohammad (saif.mohammad@nrc-cnrc.gc.ca).

Ganchev, K., Graça, J., Blitzer, J., and Taskar, B. (2012). Multi-view learning over structured and non-identical outputs. *CoRR*, **abs/1206.3256**.

Genereux, M. and Evans, R. P. (2006). Distinguishing affective states in weblogs. In *AAAI-2006 Spring Symposium on Computational Approaches to Analysing Weblogs*, pages 27–29, Stanford, California.

Gill, A. and Oberlander, J. (2002). Taking care of the linguistic features of extraversion. In *Proceedings of the Conference of the Cognitive Science Society*, pages 363–368, Sapporo, Japan.

Go, A., Bhayani, R., and Huang, L. (2009). Twitter Sentiment Classification using Distant Supervision. In *Final Projects from CS224N for Spring 2008/2009 at The Stanford Natural Language Processing Group*.

González-Ibáñez, R., Muresan, S., and Wacholder, N. (2011). Identifying sarcasm in twitter: a closer look. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 581–586, Portland, Oregon.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The WEKA data mining software: an update. *SIGKDD Explorations*, **11**, 10–18.

Hayes, N. and Joseph, S. (2003). Big 5 correlates of three measures of subjective well-being. *Personality and Individual Differences*, **34**(4), 723–727.

Holzman, L. E. and Pottenger, W. M. (2003). Classification of emotions in internet chat: An application of machine learning using speech phonemes. Technical report, Leigh University.

John, D., Boucouvalas, A. C., and Xu, Z. (2006). Representing emotional momentum within expressive internet communication. In *Proceedings of the 24th IASTED international conference on Internet and multimedia systems and applications*, pages 183–188, Anaheim, CA. ACTA Press.

John, O. P. and Srivastava, S. (1999). The big five taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin and O. P. John, editors, *Handbook of Personality Theory and Research*, pages 102–138. Guilford Press.

Karypis, G. (2003). CLUTO: A Clustering Toolkit. *Technical report, University of Minnesota*.

Kim, E., Gilbert, S., Edwards, M. J., and Graeff, E. (2009). Detecting sadness in 140 characters: Sentiment analysis of mourning Michael Jackson on Twitter. In *The Web Ecology project*.

Kosinski, M., Stillwell, D., and Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*.

Lastovicka, J. L. and Joachimsthaler, E. A. (1988). Improving the detection of personality-behavior relationships in consumer research. *Journal of Consumer Research*, **14**(4), 583–587.

Litman, D. J. and Forbes-Riley, K. (2004). Predicting student emotions in computer-human tutoring dialogues. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, ACL '04, Morristown, NJ, USA.

Liu, H., Lieberman, H., and Selker, T. (2003). A model of textual affect sensing using real-world knowledge. In *Proceedings of the 8th international conference on Intelligent user interfaces*, IUI '03, pages 125–132, New York, NY.

Ma, C., Prendinger, H., and Ishizuka, M. (2005). Emotion estimation and reasoning based on affective textual interaction. In *First International Conference on Affective Computing and Intelligent Interaction (ACII-2005)*, pages 622–628, Beijing, China.

Mairesse, F., Walker, M. A., Mehl, M. R., and Moore, R. K. (2007). Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research*, **30**(1), 457–500.

Matykiewicz, P., Duch, W., and Pestian, J. (2009). Clustering semantic spaces of suicide notes and newsgroups articles. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, BioNLP '09, pages 179–184, Stroudsburg, PA, USA.

Mihalcea, R. and Liu, H. (2006). A corpus-based approach to finding happiness. In *AAAI-2006 Spring Symposium on Computational Approaches to Analysing Weblogs*, pages 139–144. AAAI Press.

Mohammad, S. M. (2012a). From once upon a time to happily ever after: Tracking emotions in mail and books. *Decision Support Systems*, **53**(4), 730–741.

Mohammad, S. M. (2012b). Portable features for classifying emotional text. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human*

Language Technologies (NAACL-HLT 2012), Montreal, Canada. Association for Computational Linguistics.

Mohammad, S. M. and Turney, P. D. (2013). Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, **29**(3), 436–465.

Mohammad, S. M. and Yang, T. (2011). Tracking Sentiment in Mail: How Genders Differ on Emotional Axes. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2.011)*, pages 70–79, Portland, Oregon. Association for Computational Linguistics.

Myers, I. B., McCaulley, M. H., and Most, R. (1985). *Manual: A guide to the development and use of the Myers-Briggs Type Indicator*. Consulting Psychologists Press Palo Alto, CA.

Myers, L. (1962). Manual: The myers-briggs type indicator. *Educational Testing Services*.

Neviarouskaya, A., Prendinger, H., and Ishizuka, M. (2009). Compositionality principle in recognition of fine-grained emotions from text. In *Proceedings of the Proceedings of the Third International Conference on Weblogs and Social Media (ICWSM-09)*, pages 278–281, San Jose, California.

Osgood, C. E. and Walker, E. G. (1959). Motivation and language behavior: A content analysis of suicide notes. *Journal of Abnormal and Social Psychology*, **59**(1), 58–67.

Osgood, C. E., Suci, G. J., and Tannenbaum, P. (1957). *The measurement of meaning*. University of Illinois Press.

Pearl, L. and Steyvers, M. (2010). Identifying emotions, intentions, and attitudes in text using a game with a purpose. In *Proceedings of the NAACL-HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, Los Angeles, California.

Pennebaker, J. W. and King, L. A. (1999). Linguistic styles: language use as an individual difference. *Journal of Personality and Social Psychology*, **77**(6), 1296–1312.

Pestian, J. P., Matykiewicz, P., and Grupp-Phelan, J. (2008). Using natural language processing to classify suicide notes. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, BioNLP '08, pages 96–97, Stroudsburg, PA, USA. Association for Computational Linguistics.

Platt, J. (1999). Using analytic QP and sparseness to speed training of support vector machines. In *In Neural Info. Processing Systems 11*, pages 557–563. MIT Press.

Plutchik, R. (1962). *The Emotions*. New York: Random House.

Plutchik, R. (1980). A general psychoevolutionary theory of emotion. *Emotion: Theory, research, and experience*, **1**(3), 3–33.

Plutchik, R. (1985). On emotion: The chicken-and-egg problem revisited. *Motivation and Emotion*, **9**(2), 197–200.

Plutchik, R. (1994). *The psychology and biology of emotion*. New York: Harper Collins.

Qiu, L., Lin, H., Ramsay, J., and Yang, F. (2012). You are what you tweet: Personality expression and perception on twitter. *Journal of Research in Personality*.

Ravaja, N., Saari, T., Turpeinen, M., Laarni, J., Salminen, M., and Kivikangas, M. (2006). Spatial presence and emotions during video game playing: Does it matter with whom you play? *Presence: Teleoperators and Virtual Environments*, **15**(4), 381–392.

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th international joint conference on Artificial intelligence - Volume 1*, IJCAI'95, pages 448–453, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Strapparava, C. and Mihalcea, R. (2007). Semeval-2007 task 14: Affective text. In *Proceedings of SemEval-2007*, pages 70–74, Prague, Czech Republic.

Strapparava, C. and Valitutti, A. (2004). WordNet-Affect: An affective extension of WordNet. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC-2004)*, pages 1083–1086, Lisbon, Portugal.

Tett, R. P., Jackson, D. N., and Rothstein, M. (1991). Personality measures as predictors of job performance: a meta-analytic review. *Personnel Psychology*, **44**(4), 703–742.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., and Welpe, I. M. (2010). Predicting elections with twitter : What 140 characters reveal about political sentiment. *Word Journal Of The International Linguistic Association*, pages 178–185.

Turney, P. and Littman, M. L. (2003). Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems*, **21**(4).

Velásquez, J. D. (1997). Modeling emotions and other motivations in synthetic agents. In *Proceedings of the fourteenth national conference on artificial intelligence and ninth conference on Innovative applications of artificial intelligence*, AAAI'97/IAAI'97, pages 10–15. AAAI Press.

White, J. K., Hendrick, S. S., and Hendrick, C. (2004). Big five personality variables and relationship constructs. *Personality and Individual Differences*, **37**(7), 1519–1530.

Yarkoni, T. (2010). Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers. *Journal of research in personality*, **44**(3), 363–373.

Zhe, X. and Boucouvalas, A. (2002). Text-to-emotion engine for real time internet communication. In *Proceedings of the International Symposium on CSNDSP*, pages 164–168, Staffordshire University.