VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text

C.J. Hutto **Eric Gilbert**
Georgia Institute of Technology, Atlanta, GA 30032
cjhutto@gatech.edu gilbert@cc.gatech.edu

Abstract

The inherent nature of social media content poses serious challenges to practical applications of sentiment analysis. We present VADER, a simple rule-based model for general sentiment analysis, and compare its effectiveness to eleven typical state-of-practice benchmarks including LIWC, ANEW, the General Inquirer, SentiWordNet, and machine learning oriented techniques relying on Naive Bayes, Maximum Entropy, and Support Vector Machine (SVM) algorithms. Using a combination of qualitative and quantitative methods, we first construct and empirically validate a *gold-standard* list of lexical features (along with their associated sentiment intensity measures) which are specifically attuned to sentiment in microblog-like contexts. We then combine these lexical features with consideration for five general rules that embody grammatical and syntactical conventions for expressing and emphasizing sentiment intensity. Interestingly, using our parsimonious rule-based model to assess the sentiment of tweets, we find that VADER outperforms individual human raters (*F1 Classification Accuracy* = 0.96 and 0.84, respectively), and generalizes more favorably across contexts than any of our benchmarks.

1. Introduction

Sentiment analysis is useful to a wide range of problems that are of interest to human-computer interaction practitioners and researchers, as well as those from fields such as sociology, marketing and advertising, psychology, economics, and political science. The inherent nature of microblog content - such as those observed on Twitter and Facebook - poses serious challenges to practical applications of sentiment analysis. Some of these challenges stem from the sheer rate and volume of user generated social content, combined with the contextual sparseness resulting from shortness of the text and a tendency to use abbreviated language conventions to express sentiments.

A comprehensive, high quality lexicon is often essential for fast, accurate sentiment analysis on such large scales. An example of such a lexicon that has been widely used in the social media domain is the Linguistic Inquiry and Word Count (LIWC, pronounced "Luke") (Pennebaker, Francis, & Booth, 2001; Pennebaker, Chung, Ireland, Gonzales, & Booth, 2007). Sociologists, psychologists, linguists, and computer scientists find LIWC appealing because it has been extensively validated. Also, its straightforward dictionary and simple word lists are easily inspected, understood, and extended if desired. Such attributes make LIWC an attractive option to researchers looking for a reliable lexicon to extract emotional or sentiment polarity from text. Despite their pervasive use for gaging sentiment in social media contexts, these lexicons are often used with little regard for their actual suitability to the domain.

This paper describes the development, validation, and evaluation of VADER (for Valence Aware Dictionary for sEntiment Reasoning). We use a combination of qualitative and quantitative methods to produce, and then empirically validate, a *gold-standard* sentiment lexicon that is especially attuned to microblog-like contexts. We next combine these lexical features with consideration for five generalizable rules that embody grammatical and syntactical conventions that humans use when expressing or emphasizing sentiment *intensity*. We find that incorporating these heuristics improves the accuracy of the sentiment analysis engine across several domain contexts (social media text, NY Times editorials, movie reviews, and product reviews). Interestingly, the VADER lexicon performs exceptionally well in the social media domain. The correlation coefficient shows that VADER (r = 0.881) performs as well as individual human raters (r = 0.888) at matching ground truth (aggregated group mean from 20 human raters for sentiment intensity of each tweet). Surprisingly, when we further inspect the classification accuracy, we see that VADER (*F1* = 0.96) actually even outperforms individual human raters (*F1* = 0.84) at correctly classifying the sentiment of tweets into positive, neutral, or negative classes.

VADER retains (and even improves on) the benefits of traditional sentiment lexicons like LIWC: it is bigger, yet just as simply inspected, understood, quickly applied (without a need for extensive learning/training) and easily extended. Like LIWC (but unlike some other lexicons or machine learning models), the VADER sentiment lexicon is *gold-standard* quality and has been validated by humans. VADER distinguishes itself from LIWC in that it is more sensitive to sentiment expressions in social media contexts while also generalizing more favorably to other domains. We make VADER freely available for download and use.

2. Background and Related Work

Sentiment analysis, or opinion mining, is an active area of study in the field of natural language processing that analyzes people's opinions, sentiments, evaluations, attitudes, and emotions via the computational treatment of subjectivity in text. It is not our intention to review the entire body of literature concerning sentiment analysis. Indeed, such an endeavor would not be possible within the limited space available (such treatments are available in Liu (2012) and Pang & Lee (2008)). We do provide a brief overview of canonical works and techniques relevant to our study.

2.1 Sentiment Lexicons

A substantial number of sentiment analysis approaches rely greatly on an underlying sentiment (or opinion) lexicon. A *sentiment lexicon* is a list of lexical features (e.g., words) which are generally labeled according to their semantic orientation as either positive or negative (Liu, 2010). Manually creating and validating such lists of opinion-bearing features, while being among the most robust methods for generating reliable sentiment lexicons, is also one of the most time-consuming. For this reason, much of the applied research leveraging sentiment analysis relies heavily on preexisting manually constructed lexicons. Because lexicons are so useful for sentiment analysis, we briefly provide an overview of several benchmarks. We first review three widely used lexicons (LIWC[1], GI[2], Hu-Liu04[3]) in which words are categorized into binary classes (i.e., either *positive* or *negative*) according to their context free semantic orientation. We then describe three other lexicons (ANEW[4], SentiWordNet[5], and SenticNet[6]) in which words are associated with valence scores for sentiment *intensity*.

2.1.1 Semantic Orientation (Polarity-based) Lexicons

LIWC is text analysis software designed for studying the various emotional, cognitive, structural, and process components present in text samples. LIWC uses a proprietary dictionary of almost 4,500 words organized into one (or more) of 76 categories, including 905 words in two categories especially related to sentiment analysis (see Table 1):

LIWC Category	Examples	No. of Words
Positive Emotion	Love, nice, good, great	406
Negative Emotion	Hurt, ugly, sad, bad, worse	499

Table 1: Example words from two of LIWC's 76 categories. These two categories can be leveraged to construct a semantic orientation-based lexicon for sentiment analysis.

[1] www.liwc.net
[2] http://www.wjh.harvard.edu/~inquirer
[3] http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html
[4] http://csea.phhp.ufl.edu/media/anewmessage.html
[5] http://sentiwordnet.isti.cnr.it/
[6] http://sentic.net/

LIWC is well-established and has been both internally and externally validated in a process spanning more than a decade of work by psychologists, sociologists, and linguists (Pennebaker et al., 2001; Pennebaker et al., 2007). Its pedigree and validation make LIWC an attractive option to researchers looking for a reliable lexicon to extract emotional or sentiment polarity from social media text. For example, LIWC's lexicon has been used to extract indications of political sentiment from tweets (Tumasjan, Sprenger, Sandner, & Welpe, 2010), predict the onset of depression in individuals based on text from social media (De Choudhury, Gamon, Counts, & Horvitz, 2013), characterize the emotional variability of pregnant mothers from Twitter posts (De Choudhury, Counts, & Horvitz, 2013), unobtrusively measure national happiness based on Facebook status updates (Kramer, 2010), and differentiating happy romantic couples from unhappy ones based on their instant message communications (Hancock, Landrigan, & Silver, 2007). However, as Hutto, Yardi, & Gilbert (2013) point out, despite its widespread use for assessing sentiment in social media text, LIWC does not include consideration for sentiment-bearing lexical items such as acronyms, initialisms, emoticons, or slang, which are known to be important for sentiment analysis of social text (Davidov, Tsur, & Rappoport, 2010). Also, LIWC is unable to account for differences in the sentiment *intensity* of words. For example, "The food here is *exceptional*" conveys more positive intensity than "The food here is *okay*". A sentiment analysis tool using LIWC would score them equally (they each contain one positive term). Such distinctions are intuitively valuable for fine-grained sentiment analysis.

The General Inquirer (GI) is a text analysis application with one of the oldest manually constructed lexicons still in widespread use. The GI has been in development and refinement since 1966, and is designed as a tool for *content analysis*, a technique used by social scientists, political scientists, and psychologists for objectively identifying specified characteristics of messages (Stone et al., 1966). The lexicon contains more than 11K words classified into one or more of 183 categories. For our purposes, we focus on the 1,915 words labeled *Positive* and the 2,291 words labeled as *Negative*. Like LIWC, the Harvard GI lexicon has been widely used in several works to automatically determine sentiment properties of text (Esuli & Sebastiani, 2005; Kamps, Mokken, Marx, & de Rijke, 2004; Turney & Littman, 2003). However, as with LIWC, the GI suffers from a lack of coverage of sentiment-relevant lexical features common to social text, and it is ignorant of *intensity* differences among sentiment-bearing words.

Hu and Liu (Hu & Liu, 2004; Liu, Hu, & Cheng, 2005) maintain a publicly available lexicon of nearly 6,800 words (2,006 with positive semantic orientation, and 4,783 negative). Their opinion lexicon was initially constructed through a bootstrapping process (Hu & Liu, 2004) using

WordNet (Fellbaum, 1998), a well-known English lexical database in which words are clustered into groups of synonyms known as *synsets*. The Hu-Liu04 opinion lexicon has evolved over the past decade, and (unlike LIWC or the GI lexicons) is more attuned to sentiment expressions in social text and product reviews – though it still does not capture sentiment from emoticons or acronyms/initialisms.

2.1.2 Sentiment Intensity (Valence-based) Lexicons

Many applications would benefit from being able to determine not just the binary polarity (positive versus negative), but also the *strength* of the sentiment expressed in text. Just how favorably or unfavorably do people feel about a new product, movie, or legislation bill? Analysts and researchers want (and need) to be able to recognize changes in sentiment *intensity* over time in order to detect when rhetoric is heating up or cooling down (Wilson, Wiebe, & Hwa, 2004). It stands to reason that having a general lexicon with strength valences would be beneficial.

The Affective Norms for English Words (ANEW) lexicon provides a set of normative emotional ratings for 1,034 English words (Bradley & Lang, 1999). Unlike LIWC or GI, the words in ANEW have been ranked in terms of their pleasure, arousal, and dominance. ANEW words have an associated sentiment valence ranging from 1-9 (with a neutral midpoint at five), such that words with valence scores less than five are considered unpleasant/negative, and those with scores greater than five are considered pleasant/positive. For example, the valence for *betray* is 1.68, *bland* is 4.01, *dream* is 6.73, and *delight* is 8.26. These valences help researchers measure the intensity of expressed sentiment in microblogs (De Choudhury, Counts, et al., 2013; De Choudhury, Gamon, et al., 2013; Nielsen, 2011) – an important dimension beyond simple binary orientations of positive and negative. Nevertheless, as with LIWC and GI, the ANEW lexicon is also insensitive to common sentiment-relevant lexical features in social text.

SentiWordNet is an extension of WordNet (Fellbaum, 1998) in which 147,306 synsets are annotated with three numerical scores relating to positivity, negativity, and objectivity (neutrality) (Baccianella, Esuli, & Sebastiani, 2010). Each score ranges from 0.0 to 1.0, and their sum is 1.0 for each synset. The scores were calculated using a complex mix of semi-supervised algorithms (propagation methods and classifiers). It is thus not a *gold standard* resource like WordNet, LIWC, GI, or ANEW (which were all 100% curated by humans), but it is useful for a wide range of tasks. We interface with SentiWordNet via Python's Natural Language Toolkit[7] (NLTK), and use the difference of each sysnset's positive and negative scores as its sentiment *valence* to distinguish differences in the sentiment intensity of words. The SentiWordNet lexicon is very noisy; a large majority of synsets have no positive or negative polarity. It also fails to account for sentiment-bearing lexical features relevant to text in microblogs.

SenticNet is a publicly available semantic and affective resource for concept-level opinion and sentiment analysis (Cambria, Havasi, & Hussain, 2012). SenticNet is constructed by means of *sentic computing*, a paradigm that exploits both AI and Semantic Web techniques to process natural language opinions via an ensemble of graph-mining and dimensionality-reduction techniques (Cambria, Speer, Havasi, & Hussain, 2010). The SenticNet lexicon consists of 14,244 common sense concepts such as *wrath*, *adoration*, *woe*, and *admiration* with information associated with (among other things) the concept's sentiment *polarity*, a numeric value on a continuous scale ranging from –1 to 1. We access the SenticNet polarity score using the online SenticNet API and a publicly available Python package[8].

2.1.3 Lexicons and Context-Awareness

Whether one is using binary polarity-based lexicons or more nuanced valence-based lexicons, it is possible to improve sentiment analysis performance by understanding deeper lexical properties (e.g., parts-of-speech) for more context awareness. For example, a lexicon may be further tuned according to a process of word-sense disambiguation (WSD) (Akkaya, Wiebe, & Mihalcea, 2009). Word-sense disambiguation refers to the process of identifying which sense of a word is used in a sentence when the word has multiple meanings (i.e. its contextual meaning). For example, using WSD, we can distinguish that the word *catch* has negative sentiment in "At first glance the contract looks good, but there's a *catch*", but is neutral in "The fisherman plans to sell his *catch* at the market". We use a publicly available Python package[9] that performs sentiment classification with word-sense disambiguation.

Despite their ubiquity for evaluating sentiment in social media contexts, there are generally three shortcomings of lexicon-based sentiment analysis approaches: 1) they have trouble with coverage, often ignoring important lexical features which are especially relevant to social text in microblogs, 2) some lexicons ignore general sentiment intensity differentials for features within the lexicon, and 3) acquiring a new set of (human validated gold standard) lexical features – along with their associated sentiment valence scores – can be a very time consuming and labor intensive process. We view the current study as an opportunity not only to address this gap by constructing just such a lexicon and providing it to the broader research community, but also a chance to compare its efficacy against other well-established lexicons with regards to sentiment analysis of social media text and other domains.

[7] http://www.nltk.org

[8] senticnet 0.3.2 (https://pypi.python.org/pypi/senticnet)

[9] https://pypi.python.org/pypi/sentiment_classifier/0.5

2.2 Machine Learning Approaches

Because manually creating and validating a comprehensive sentiment lexicon is labor and time intensive, much work has explored automated means of identifying sentiment-relevant features in text. Typical state of the art practices incorporate machine learning approaches to "learn" the sentiment-relevant features of text.

The Naive Bayes (NB) classifier is a simple classifier that relies on Bayesian probability and the naive assumption that feature probabilities are independent of one another. Maximum Entropy (MaxEnt, or ME) is a general purpose machine learning technique belonging to the class of exponential models using multinomial logistic regression. Unlike NB, ME makes no conditional independence assumption between features, and thereby accounts for information entropy (feature weightings). Support Vector Machines (SVMs) differ from both NB and ME models in that SVMs are non-probability classifiers which operate by separating data points in space using one or more *hyperplanes* (centerlines of the gaps separating different classes). We use the Python-based machine learning algorithms from scikit-learn.org for the NB, ME, SVM-Classification (SVM-C) and SVM-Regression (SVM-R) models.

Machine learning approaches are not without drawbacks. First, they require (often extensive) training data which are, as with validated sentiment lexicons, sometimes troublesome to acquire. Second, they depend on the training set to represent as many features as possible (which often, they do not – especially in the case of the short, sparse text of social media). Third, they are often more computationally expensive in terms of CPU processing, memory requirements, and training/classification time (which restricts the ability to assess sentiment on streaming data). Fourth, they often derive features "behind the scenes" inside of a black box that is not (easily) human-interpretable and are therefore more difficult to either generalize, modify, or extend (e.g., to other domains).

3. Methods

Our approach seeks to leverage the advantages of parsimonious rule-based modeling to construct a computational sentiment analysis engine that 1) works well on social media style text, yet readily generalizes to multiple domains, 2) requires no training data, but is constructed from a generalizable, valence-based, human-curated gold standard sentiment lexicon 3) is fast enough to be used online with streaming data, and 4) does not severely suffer from a speed-performance tradeoff.

Figure 1 provides an overview of the research process and summarizes the methods used in this study. In essence, this paper reports on three interrelated efforts: 1) the development and validation of a gold standard sentiment lexicon that is sensitive both the *polarity* and the *intensity* of sentiments expressed in social media microblogs (but which is also generally applicable to sentiment analysis in other domains); 2) the identification and subsequent experimental evaluation of generalizable rules regarding conventional uses of grammatical and syntactical aspects of text for assessing sentiment intensity; and 3) comparing the performance of a parsimonious lexicon and rule-based model against other established and/or typical sentiment analysis baselines. In each of these three efforts, we incorporate an explicit human-centric approach. Specifically, we combine qualitative analysis with empirical validation and experimental investigations leveraging the wisdom-of-the-crowd (Surowiecki, 2004).

3.1 Constructing and Validating a Valence-Aware Sentiment Lexicon: A Human-Centered Approach

Manually creating (much less, validating) a comprehensive sentiment lexicon is a labor intensive and sometimes error prone process, so it is no wonder that many opinion mining researchers and practitioners rely so heavily on existing lexicons as primary resources. There is, of course, a great

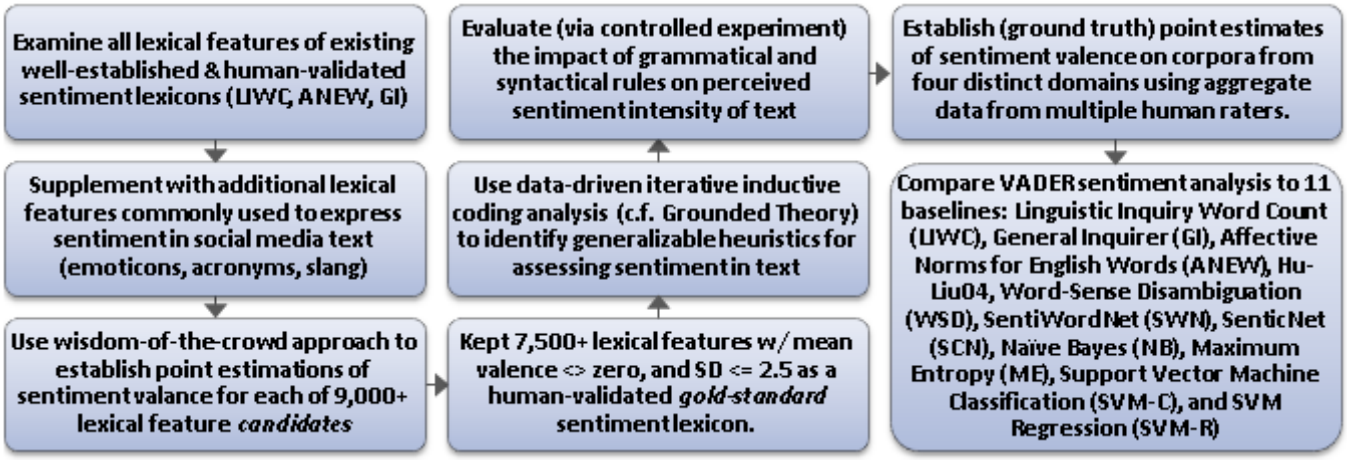

Figure 1: Methods and process approach overview.

deal of overlap in the vocabulary covered by such lexicons; however, there are also numerous items unique to each.

We begin by constructing a list inspired by examining existing well-established sentiment word-banks (LIWC, ANEW, and GI). To this, we next incorporate numerous lexical features common to sentiment expression in microblogs, including a full list of Western-style emoticons[10] (for example, ":-)" denotes a "smiley face" and generally indicates positive sentiment), sentiment-related acronyms and initialisms[11] (e.g., LOL and WTF are both sentiment-laden initialisms), and commonly used slang[12] with sentiment value (e.g., "nah", "meh" and "giggly"). This process provided us with over 9,000 lexical feature *candidates*.

Next, we assessed the general applicability of each feature candidate to sentiment expressions. We used a wisdom-of-the-crowd[13] (WotC) approach (Surowiecki, 2004) to acquire a valid point estimate for the sentiment valence (intensity) of each context-free candidate feature. We collected intensity ratings on each of our candidate lexical features from ten independent human raters (for a total of 90,000+ ratings). Features were rated on a scale from "[–4] Extremely Negative" to "[4] Extremely Positive", with allowance for "[0] Neutral (or Neither, N/A)". Ratings were obtained using Amazon Mechanical Turk (AMT), a micro-labor website where workers perform minor tasks in exchange for a small amount of money (see subsection 3.1.1 for details on how we were able to consistently obtain high quality, generalizable results from AMT workers). Figure 2 illustrates the user interface implemented for acquiring valid point estimates of sentiment intensity for each context-free candidate feature comprising the VADER sentiment lexicon. (A similar UI was leveraged for all of the evaluation and validation activities described in subsections 3.1, 3.2, 3.3, and 3.4.) We kept every lexical feature that had a non-zero mean rating, and whose standard deviation was less than 2.5 as determined by the aggregate of ten independent raters. This left us with just over 7,500 lexical features with validated valence scores that indicated both the sentiment *polarity* (positive/negative), and the sentiment *intensity* on a scale from –4 to +4. For example, the word "okay" has a positive valence of 0.9, "good" is 1.9, and "great" is 3.1, whereas "horrible" is –2.5, the frowning emoticon ":(" is –2.2, and "sucks" and "sux" are both –1.5. This gold standard list of features, with associated valence for each feature, comprises VADER's sentiment lexicon, and is available for download from our website[14].

[10] http://en.wikipedia.org/wiki/List_of_emoticons#Western
[11] http://en.wikipedia.org/wiki/List_of_acronyms
[12] http://www.internetslang.com/
[13] *Wisdom-of-the-crowd* is the process of incorporating aggregated opinions from a collection of individuals to answer a question. The process has been found to be as good as (often better than) estimates from lone individuals, even experts.
[14] http://comp.social.gatech.edu/papers/

3.1.1 Screening, Training, Selecting, and Data Quality Checking Crowd-Sourced Evaluations and Validations

Previous linguistic rating experiments using a WotC approach on AMT have shown to be reliable – sometimes even outperforming expert raters (Snow, O'Connor, Jurafsky, & Ng, 2008). On the other hand, prior work has also advised on methods to reduce the amount of noise from AMT workers who may produce poor quality work (Downs, Holbrook, Sheng, & Cranor, 2010; Kittur, Chi, & Suh, 2008). We therefore implemented four quality control processes to help ensure we received meaningful data from our AMT raters.

First, every rater was prescreened for English language reading comprehension – each rater had to individually score an 80% or higher on a standardized college-level reading comprehension test.

Second, every prescreened rater then had to complete an online sentiment rating training and orientation session, and score 90% or higher for matching the known (pre-validated) mean sentiment rating of lexical items which included individual words, emoticons, acronyms, sentences, tweets, and text snippets (e.g., sentence segments, or phrases). The user interface employed during the sentiment training (Figure 2) always matched the specific sentiment rating tasks discussed in this paper. The training helped to ensure consistency in the rating rubric used by each independent rater.

Third, every batch of 25 features contained five "golden items" with a known (pre-validated) sentiment rating distribution. If a worker was more than one standard deviation away from the mean of this known distribution on three or more of the five golden items, we discarded all 25 ratings in the batch from this worker.

Finally, we implemented a bonus program to incentivize and reward the highest quality work. For example, we asked workers to select the valence score that they thought "*most other people*" would choose for the given lexical feature (early/iterative pilot testing revealed that wording the instructions in this manner garnered a much tighter standard deviation without significantly affecting the mean sentiment rating, allowing us to achieve higher quality (generalized) results while being more economical).

We compensated AMT workers $0.25 for each batch of 25 items they rated, with an additional $0.25 incentive bonus for all workers who successfully matched the group mean (within 1.5 standard deviations) on at least 20 of 25 responses in each batch. Using these four quality control methods, we achieved remarkable value in the data obtained from our AMT workers – we paid incentive bonuses for high quality to at least 90% of raters for most batches.

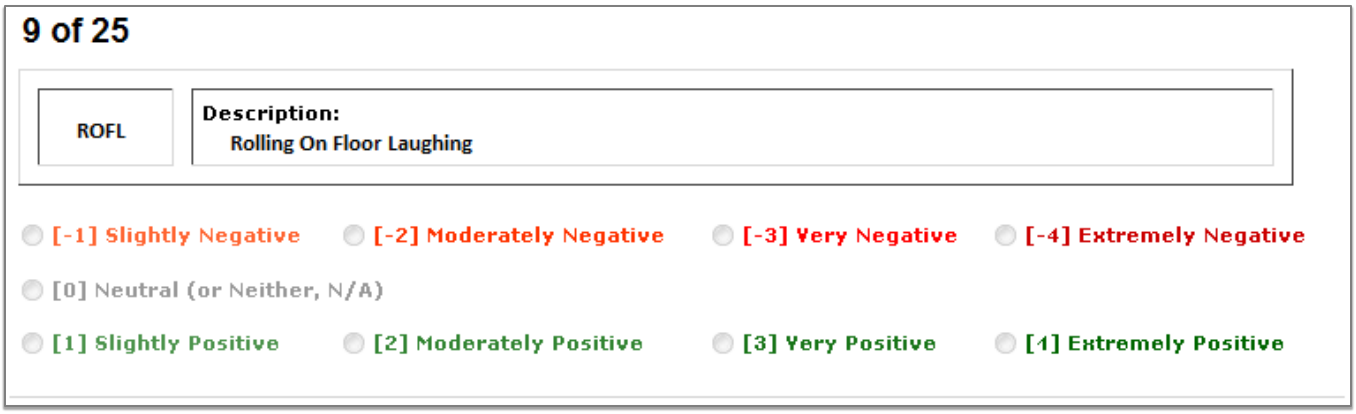

Figure 2: Example of the interface implemented for acquiring valid point estimates of sentiment valence (intensity) for each context-free candidate feature comprising the VADER sentiment lexicon. A similar UI was used for all rating activities described in sections 3.1-3.4.

3.2 Identifying Generalizable Heuristics Humans Use to Assess Sentiment Intensity in Text

We next analyze a purposeful sample of 400 positive and 400 negative social media text snippets (tweets). We selected this sample from a larger initial set of 10K random tweets pulled from Twitter's public timeline based on their sentiment scores using the Pattern.en sentiment analysis engine[15] (they were the top 400 most positive and negative tweets in the set). Pattern is a web mining module for Python, and the Pattern.en module is a natural language processing (NLP) toolkit (De Smedt & Daelemans, 2012) that leverages WordNet to score sentiment according to the English adjectives used in the text.

Next, two human experts individually scrutinized all 800 tweets, and independently scored their sentiment intensity on a scale from –4 to +4. Following a data-driven inductive coding technique similar to the Grounded Theory approach (Strauss & Corbin, 1998), we next used qualitative analysis techniques to identify properties and characteristics of the text which affect the perceived sentiment *intensity* of the text. This deep qualitative analysis resulted in isolating five generalizable heuristics based on grammatical and syntactical cues to convey changes to sentiment intensity. Importantly, these heuristics go beyond what would normally be captured in a typical bag-of-words model. They incorporate word-order sensitive relationships between terms:

1. Punctuation, namely the exclamation point (!), increases the magnitude of the intensity without modifying the semantic orientation. For example, "*The food here is good!!!*" is more intense than "*The food here is good.*"
2. Capitalization, specifically using ALL-CAPS to emphasize a sentiment-relevant word in the presence of other non-capitalized words, increases the magnitude of the sentiment intensity without affecting the semantic orientation. For example, "*The food here is GREAT!*" conveys more intensity than "*The food here is great!*"
3. Degree modifiers (also called *intensifiers*, *booster words*, or *degree adverbs*) impact sentiment intensity by either increasing or decreasing the intensity. For example, "*The service here is extremely good*" is more intense than "*The service here is good*", whereas "*The service here is marginally good*" reduces the intensity.
4. The contrastive conjunction "*but*" signals a shift in sentiment polarity, with the sentiment of the text following the conjunction being dominant. "*The food here is great, but the service is horrible*" has mixed sentiment, with the latter half dictating the overall rating.
5. By examining the tri-gram preceding a sentiment-laden lexical feature, we catch nearly 90% of cases where negation flips the polarity of the text. A negated sentence would be "*The food here isn't really all that great*".

3.3 Controlled Experiments to Evaluate Impact of Grammatical and Syntactical Heuristics

Using the general heuristics we just identified, we next selected 30 baseline tweets and manufactured six to ten variations of the exact same text, controlling the specific grammatical or syntactical feature that is presented as an independent variable in a small experiment. With all of the variations, we end up with 200 contrived tweets, which we then randomly insert into a new set of 800 tweets similar to those used during our qualitative analysis. We next asked 30 independent AMT workers to rate the sentiment intensity of all 1000 tweets to assess the impact of these features on perceived sentiment intensity. (AMT workers were all screened, trained, and data quality checked as described in subsection 3.1.1). Table 2 illustrates some examples of contrived variations on a given baseline:

[15] http://www.clips.ua.ac.be/pages/pattern-en#sentiment

Test Condition	Example Text
Baseline	Yay. Another good phone interview.
Punctuation1	Yay! Another good phone interview!
Punctuation1 + Degree Mod.	Yay! Another extremely good phone interview!
Punctuation2	Yay!! Another good phone interview!!
Capitalization	YAY. Another GOOD phone interview.
Punct1 + Cap.	YAY! Another GOOD phone interview!
Punct2 + Cap.	YAY!! Another GOOD phone interview!!
Punct3 + Cap.	YAY!!! Another GOOD phone interview!!!
Punct3 + Cap. + Degree Mod.	YAY!!! Another EXTREMELY GOOD phone interview!!!

Table 2: Example of baseline text with eight test conditions comprised of grammatical and syntactical variations.

Table 3 shows the *t*-test statistic, *p*-value, mean of differences for rank ordered data points between each distribution, and 95% confidence intervals:

Test Condition	*t*	*p*	Diff.	95% C.I.
Punctuation (. vs !)	19.02	< 2.2e-16	0.291	0.261 - 0.322
Punctuation (! vs !!)	16.53	2.7e-16	0.215	0.188 - 0.241
Punctuation (!! vs !!!)	14.07	1.7e-14	0.208	0.178 - 0.239
All CAPS (w/o vs w)	28.95	< 2.2e-16	0.733	0.682 - 0.784
Deg. Mod. (w/o vs w)	9.01	6.7e-10	0.293	0.227 - 0.360

Table 3: Statistics associated with grammatical and syntactical cues for expressing sentiment intensity. Differences in means were all statistically significant beyond the 0.001 level.

We incorporated the mean differences between each distribution into VADER's rule-based model. For example, from Table 3, we see that for 95% of the data, using an exclamation point (relative to a period or no punctuation at all) increased the intensity by 0.261 to 0.322, with a mean difference of 0.291 on a rating scale from 1 to 4 (we use absolute value scale here for simplicity, because it did not matter whether the text was positive or negative, using an exclamation made it equally more extreme in either case). We incorporated consideration for rule 4 by splitting the text into segments around the contrastive conjunction "*but*", and diminished the total sentiment intensity of the text preceding the conjunction by 50% while increasing the sentiment intensity of the post-conjunction text by 50%.

3.4 Ground Truth in Multiple Domain Contexts

We next obtained gold standard (human-validated) ground truth regarding sentiment intensity on corpora representing four distinct domain contexts. For this purpose, we recruited 20 independent human raters from AMT (raters were all screened, trained, and data quality checked consistent with the process described in subsection 3.1.1 and Figure 2). All four sentiment-intensity annotated corpora are available for download from our website[14]:

1. Social media text: includes 4,000 tweets pulled from Twitter's public timeline (with varied times and days of posting), plus 200 contrived tweets that specifically test syntactical and grammatical conventions of conveying differences in sentiment intensity.
2. Movie reviews: includes 10,605 sentence-level snippets from rotten.tomatoes.com. The snippets were derived from an original set of 2000 movie reviews (1000 positive and 1000 negative) in Pang & Lee (2004); we used the NLTK tokenizer to segment the reviews into sentence phrases, and added sentiment intensity ratings.
3. Technical product reviews: includes 3,708 sentence-level snippets from 309 customer reviews on 5 different products. The reviews were originally used in Hu & Liu (2004); we added sentiment intensity ratings.
4. Opinion news articles: includes 5,190 sentence-level snippets from 500 New York Times opinion editorials.

4. Results

In order to evaluate our results directly against the broader body of literature, we assess both a) the correlation of computed raw sentiment intensity rating to gold standard ground truth, i.e., the mean sentiment rating from 20 prescreened and appropriately trained human raters, as well as b) the multiclass (positive, negative, neutral) classification metrics of *precision*, *recall*, and *F1 score*. In statistical analysis of classifier performance, *precision* is the number of true classifications (i.e. the number of items labeled as a particular class that match the known gold standard classification) divided by the total number of elements labeled as that class (including both correct and incorrect classifications). *Recall* is the number of true classifications divided by the total number of elements that are known to belong to the class; low recall is an indication that known elements of a class were missed. The *F1 score* is the harmonic mean of precision and recall, and represents the overall accuracy.

We compared the VADER sentiment lexicon to seven other well-established sentiment analysis lexicons: Linguistic Inquiry Word Count (LIWC), General Inquirer (GI), Affective Norms for English Words (ANEW), SentiWordNet (SWN), SenticNet (SCN), Word-Sense Disambiguation (WSD) using WordNet, and the Hu-Liu04 opinion lexicon. For fairness to each lexicon, *all comparisons utilized VADER's rule-based model for processing syntactical and grammatical cues* – the only difference were the features represented within the actual lexicons themselves. As Figure 3 and Table 4 both show, the VADER lexicon performs exceptionally well in the social media domain, and generalizes favorably. The Pearson Product Moment Correlation Coefficient shows that VADER (r = 0.881) performs as well as individual human raters (r = 0.888) at matching ground truth (aggregated group mean from 20 human raters for sentiment intensity of each tweet). Surprisingly, when we further inspect the classification

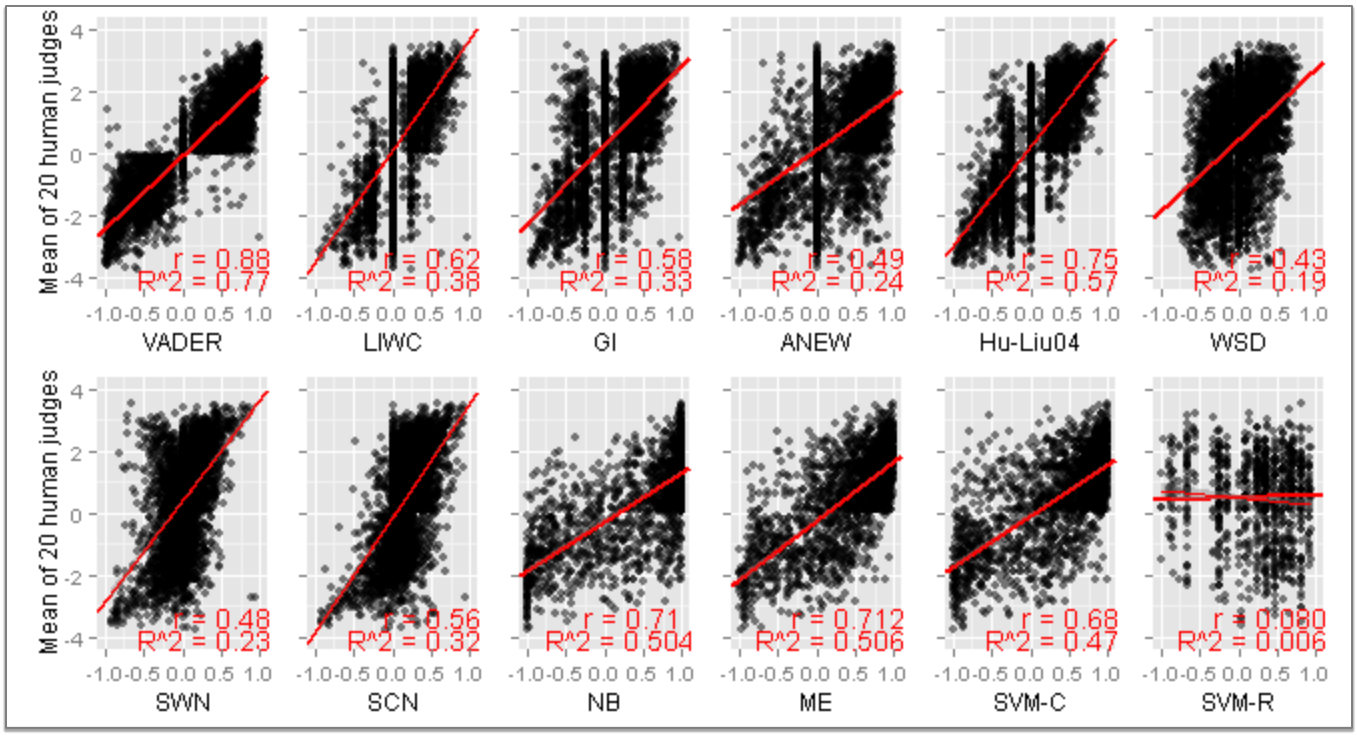

Figure 3: Sentiment scores from VADER and 11 other highly regarded sentiment analysis tools/techniques on a corpus of over 4K tweets. Although this figure specifically portrays correlation, it also helps to visually depict (and contrast) VADER's classification precision, recall, and F1 accuracy within this domain (see Table 4). Each subplot can be roughly considered as having four quadrants: true negatives (lower left), true positives (upper right), false negatives (upper left), and false positives (lower right).

	Correlation to ground truth (mean of 20 human raters)	3-class (positive, negative, neutral) Classification Accuracy Metrics			Ordinal Rank (by F1)		Correlation to ground truth (mean of 20 human raters)	3-class (positive, negative, neutral) Classification Accuracy Metrics		
		Overall Precision	Overall Recall	Overall F1 score				Overall Precision	Overall Recall	Overall F1 score
Social Media Text (4,200 Tweets)							**Movie Reviews (10,605 review snippets)**			
Ind. Humans	**0.888**	0.95	0.76	0.84	2	1	0.899	0.95	0.90	0.92
VADER	0.881	**0.99**	**0.94**	**0.96**	1*	2	0.451	0.70	0.55	0.61
Hu-Liu04	0.756	0.94	0.66	0.77	3	3	0.416	0.66	0.56	0.59
SCN	0.568	0.81	0.75	0.75	4	7	0.210	0.60	0.53	0.44
GI	0.580	0.84	0.58	0.69	5	5	0.343	0.66	0.50	0.55
SWN	0.488	0.75	0.62	0.67	6	4	0.251	0.60	0.55	0.57
LIWC	0.622	0.94	0.48	0.63	7	9	0.152	0.61	0.22	0.31
ANEW	0.492	0.83	0.48	0.60	8	8	0.156	0.57	0.36	0.40
WSD	0.438	0.70	0.49	0.56	9	6	0.349	0.58	0.50	0.52
Amazon.com Product Reviews (3,708 review snippets)							**NY Times Editorials (5,190 article snippets)**			
Ind. Humans	0.911	0.94	0.80	0.85	1	1	0.745	0.87	0.55	0.65
VADER	0.565	0.78	0.55	0.63	2	2	0.492	0.69	0.49	0.55
Hu-Liu04	0.571	0.74	0.56	0.62	3	3	0.487	0.70	0.45	0.52
SCN	0.316	0.64	0.60	0.51	7	7	0.252	0.62	0.47	0.38
GI	0.385	0.67	0.49	0.55	5	5	0.362	0.65	0.44	0.49
SWN	0.325	0.61	0.54	0.57	4	4	0.262	0.57	0.49	0.52
LIWC	0.313	0.73	0.29	0.36	9	9	0.220	0.66	0.17	0.21
ANEW	0.257	0.69	0.33	0.39	8	8	0.202	0.59	0.32	0.35
WSD	0.324	0.60	0.51	0.55	6	6	0.218	0.55	0.45	0.47

Table 4: VADER 3-class classification performance as compared to individual human raters and 7 established lexicon baselines across four distinct domain contexts (clockwise from upper left: tweets, movie reviews, product reviews, opinion news articles).

accuracy (with classification thresholds set at –0.05 and +0.05 for all normalized sentiment scores between -1 and 1), we can see that VADER (*F1* = 0.96) actually outperforms even individual human raters (*F1* = 0.84) at correctly classifying the sentiment of tweets. Notice how the LIWC, GI, ANEW, and Hu-liu04 results in Figure 3 show a concentration of tweets incorrectly classified as neutral. Presumably, this is due to lack of coverage for the sentiment-oriented language of social media text, which is often expressed using emoticons, slang, or abbreviated text such as acronyms and initialisms.

The lexicons for the machine learning algorithms were all constructed by training those models on half the data (again, incorporating all rules), with the other half being held out for testing. While some algorithms performed decently on test data from the specific domain for which it was expressly trained, *they do not significantly outstrip the simple model we use*. Indeed, in three out of four cases, VADER performs as well or better *across* domains than the machine learning approaches do in the *same* domain for which they were trained. Table 5 explicitly shows this, and also highlights another advantage of VADER – its simplicity makes it computationally efficient, unlike some SVM models, which were unable to fully process the data from the larger corpora (movie reviews and NYT editorials) even on a multicore system with large RAM:

	3-Class Classification Accuracy (F1 scores) Test Sets			
	Tweets	Movie	Amazon	NYT
VADER	**0.96**	0.61	**0.63**	**0.55**
NB (tweets)	0.84	0.53	0.53	0.42
ME (tweets)	0.83	0.56	0.58	0.45
SVM-C (tweets)	0.83	0.56	0.55	0.46
SVM-R (tweets)	0.65	0.49	0.51	0.46
NB (movie)	0.56	**0.75**	0.49	0.44
ME (movie)	0.56	**0.75**	0.51	0.45
NB (amazon)	0.69	0.55	0.61	0.48
ME (amazon)	0.67	0.55	0.60	0.43
SVM-C (amazon)	0.64	0.55	0.58	0.42
SVM-R (amazon)	0.54	0.49	0.48	0.44
NB (nyt)	0.59	0.56	0.51	0.49
ME (nyt)	0.58	0.55	0.51	0.50

Table 5: Three-class accuracy (F1 scores) for each machine trained model (and the corpus it was trained on) as tested against every other domain context (SVM models for the movie and NYT data were too intensive for our multicore CPUs with 94GB RAM)

As discussed in subsections 3.2 and 3.3, we identified and quantified the impact of several generalizable heuristics that humans use when distinguishing between degrees of sentiment intensity. By incorporating these heuristics into VADER's rule-based model, we drastically improved both the correlation to ground truth as well as the classification accuracy of the sentiment analysis engine. Importantly, these improvements are realized *independent of the lexicon or ML model that was used*. That is, when we fairly apply the rules to all lexicons and ML algorithms, we achieve better correlation coefficients (mean *r* increase of 5.2%) and better accuracies (mean F1 increase of 2.1%). Consistent with prior work (Agarwal, Xie, Vovsha, Rambow, & Passonneau, 2011; Davidov et al., 2010; Shastri, Parvathy, Kumar, Wesley, & Balakrishnan, 2010), we find that grammatical features (conventions of use for punctuation and capitalization) and consideration for degree modifiers like "very" or "extremely" prove to be useful cues for distinguishing differences in sentiment intensity. Other syntactical considerations identified via qualitative analysis (negation, degree modifiers, and contrastive conjunctions) also help make VADER successful, and is consistent with prior work (Agarwal et al., 2011; Ding, Liu, & Yu, 2008; Lu, Castellanos, Dayal, & Zhai, 2011; Socher et al., 2013).

5. Discussion

Recent work by Socher et. al (2013) does an excellent job of summarizing (and pushing) the current state of the art for fine-grained sentence-level sentiment analysis by supervised machine learning models. As part of their excellent work using recursive deep models for assessing semantic compositionality over a sentiment tree bank, they report that the state-of-the-art regarding accuracy for simple binary (positive/negative) classification on single sentences is around 80%, and that for the more difficult multiclass case that includes a third (neutral) class, accuracies tend to hover in the 60% range for social media text (c.f. Agarwal et. al, (2011); Wang et. al (2012)). We find it very encouraging, therefore, to report that the results from VADER's simple rule-based approach are on par with such sophisticated benchmarks. However, when compared to sophisticated machine learning techniques, the simplicity of VADER carries several advantages. First, it is both quick and computationally economical *without sacrificing accuracy*. Running directly from a standard modern laptop computer with typical, moderate specifications (e.g., 3GHz processor and 6GB RAM), a corpus that takes a fraction of a second to analyze with VADER can take hours when using more complex models like SVM (if training is required) or tens of minutes if the model has been previously trained. Second, the lexicon and rules used by VADER are directly accessible, not hidden within a machine-access-only black-box. VADER is therefore easily inspected, understood, extended or modified. By exposing both the lexicon and rule-based model, VADER makes the inner workings of the sentiment analysis engine more accessible (and thus, more interpretable) to a broader human audience beyond the computer science community. Sociologists, psychologists, marketing researchers, or linguists who are comfortable using LIWC should also be able to use VADER. Third, by utilizing a *general* (human-validated) sentiment lexicon and *general* rules related to grammar and

syntax, VADER is at once both self-contained and domain agnostic – it does not require an extensive set of training data, yet it performs well in diverse domains. We stress that in no way do we intend to convey that complex or sophisticated techniques are in any wrong or bad. Instead we show that a simple, human-centric, interpretable, computationally efficient approach can produce high quality results – even outperforming individual human raters.

6. Conclusion

We report the systematic development and evaluation of VADER (Valence Aware Dictionary for sEntiment Reasoning). Using a combination of qualitative and quantitative methods, we construct and empirically validate a *gold-standard* list of lexical features (along with their associated sentiment intensity measures) which are specifically attuned to sentiment in microblog-like contexts. We then combine these lexical features with consideration for five general rules that embody grammatical and syntactical conventions for expressing and emphasizing sentiment intensity. The results are not only encouraging – they are indeed quite remarkable; VADER performed as well as (and in most cases, *better than*) eleven other highly regarded sentiment analysis tools. Our results highlight the gains to be made in computer science when the human is incorporated as a central part of the development process.

7. References

Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. (2011). Sentiment analysis of Twitter data. In *Proc. WLSM-11*s.

Akkaya, C., Wiebe, J., & Mihalcea, R. (2009). Subjectivity word sense disambiguation. In *Proc. EMNLP-09*.

Baccianella, S., Esuli, A., & Sebastiani, F. (2010). SentiWordNet 3.0. In *Proc. of LREC-10.*

Bradley, M. M., & Lang, P. J. (1999). *Affective norms for English words (ANEW): Instruction manual and affective ratings*.

Cambria, E., Havasi, C., & Hussain, A. (2012). SenticNet 2. In *Proc. AAAI IFAI RSC-12*.

Cambria, E., Speer, R., Havasi, C., & Hussain, A. (2010). SenticNet. In *Proc. of AAAI SCK-10*.

Davidov, D., Tsur, O., & Rappoport, A. (2010). Enhanced Sentiment Learning Using Twitter Hashtags and Smileys. *ICCL-10*.

De Choudhury, M., Counts, S., & Horvitz, E. (2013). Predicting Postpartum Changes in Emotion and Behavior via Social Media. In *Proc. CHI-13*.

De Choudhury, M., Gamon, M., Counts, S., & Horvitz, E. (2013). Predicting Depression via Social Media. In *Proc. ICWSM-13*.

De Smedt, T., & Daelemans, W. (2012). Pattern for Python. *Journal of Machine Learning Research*, *13*, 2063–2067.

Ding, X., Liu, B., & Yu, P. S. (2008). A holistic lexicon-based approach to opinion mining. In *Proc. ICWSDM-08*.

Downs, J. S., Holbrook, M. B., Sheng, S., & Cranor, L. F. (2010). Are your participants gaming the system?. In *Proc. CHI-10*.

Esuli, A., & Sebastiani, F. (2005). Determining the semantic orientation of terms through gloss classification. In *Proc. ICIKM-05*.

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Hancock, J. T., Landrigan, C., & Silver, C. (2007). Expressing emotion in text-based communication. In *Proc. CHI-07*.

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. In *Proc. SIGKDD KDM-04*.

Hutto, C. J., Yardi, S., & Gilbert, E. (2013). A Longitudinal Study of Follow Predictors on Twitter. In *Proc. CHI-13*.

Kamps, J., Mokken, R. J., Marx, M., & de Rijke, M. (2004). Using WordNet to measure semantic orientation. In *Proc. LREC-04*.

Kittur, A., Chi, E. H., & Suh, B. (2008). Crowdsourcing user studies with Mechanical Turk. In *Proc. CHI-08*.

Kramer, A. (2010). An unobtrusive behavioral model of "gross national happiness." In *Proc. CHI-10*.

Liu, B. (2010). Sentiment Analysis and Subjectivity. In N. Indurkhya & F. Damerau (Eds.), *Handbook of Natural Language Processing* (2nd ed.). Boca Raton, FL: Chapman & Hall.

Liu, B. (2012). *Sentiment Analysis and Opinion Mining*. San Rafael, CA: Morgan & Claypool.

Liu, B., Hu, M., & Cheng, J. (2005). Opinion Observer: Analyzing and Comparing Opinions on the Web. In *Proc. WWW-05*.

Lu, Y., Castellanos, M., Dayal, U., & Zhai, C. (2011). Automatic construction of a context-aware sentiment lexicon: an optimization approach. In *Proc. WWW-11*.

Nielsen, F. A. (2011). A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. In *Proc. ESWC-11*.

Pang, B., & Lee, L. (2004). A sentimental education: sentiment analysis using subjectivity summarization. In *Proc. ACL-04*.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations & Trends in Information Retrieval*, *2*(1), 1–135.

Pennebaker, J. W., Chung, C. K., Ireland, M., Gonzales, A., & Booth, R. J. (2007). *The development and psychometric properties of LIWC2007*. Austin, TX: LIWC.net.

Pennebaker, J. W., Francis, M., & Booth, R. (2001). *Linguistic Inquiry and Word Count: LIWC 2001*. Mahwah, NJ: Erlbaum.

Shastri, L., Parvathy, A. G., Kumar, A., Wesley, J., & Balakrishnan, R. (2010). Sentiment Extraction. *IAAI-10*.

Snow, R., O'Connor, B., Jurafsky, D., & Ng, A. Y. (2008). Cheap and Fast - But is it Good?. In *Proc. EMNLP-08*.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C., Ng, A., & Potts, C. (2013). Recursive Deep Models for Semantic Compositionality Over Sentiment Treebank. In *Proc. EMNLP-13*.

Stone, P. J., Dunphy, D. C., Smith, M. S., & Ogilvie, D. M. (1966). *General Inquirer*. Cambridge, MA: MIT Press.

Strauss, A. L., & Corbin, J. (1998). *Basics of Qualitative Research*. Thousand Oaks, CA: Sage Publications.

Surowiecki, J. (2004). *The Wisdom of Crowds*. NY, NY: Anchor.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Predicting Elections with Twitter:. In *Proc. ICWSM-10*.

Turney, P. D., & Littman, M. L. (2003). Measuring praise and criticism. *ACM Trans. Inf. Syst.*, *21*(4), 315–346.

Wang, H., Can, D., Kazemzadeh, A., Bar, F., & Narayanan, S. (2012). A system…real-time Twitter sentiment analysis. *ACL-12*.

Wilson, T., Wiebe, J., & Hwa, R. (2004). Just how mad are you? finding strong and weak opinion clauses. In *Proc. NCAI-04*s.