Choosing the Right Words: Characterizing and Reducing Error of the Word Count Approach

H. Andrew Schwartz,[1] **Johannes Eichstaedt,**[1] **Lukasz Dziurzynski,**[1] **Eduardo Blanco,**[2]
Margaret L. Kern,[1] **Stephanie Ramones,**[1] **Martin Seligman,**[1] **and Lyle Ungar**[1]
[1]University of Pennsylvania
[2]Lymba Corporation
hansens@seas.upenn.edu

Abstract

Social scientists are increasingly using the vast amount of text available on social media to measure variation in happiness and other psychological states. Such studies count words deemed to be indicators of happiness and track how the word frequencies change across locations or time. This *word count* approach is simple and scalable, yet often picks up false signals, as words can appear in different contexts and take on different meanings. We characterize the types of errors that occur using the word count approach, and find lexical ambiguity to be the most prevalent. We then show that one can reduce error with a simple refinement to such lexica by automatically eliminating highly ambiguous words. The resulting refined lexica improve precision as measured by human judgments of word occurrences in Facebook posts.

1 Introduction

Massive social media corpora, such as blogs, tweets, and Facebook statuses have recently peaked the interest of social scientists. Compared to traditional samples in tens or hundreds, social media sample sizes are orders of magnitude larger, often containing millions or billions of posts or queries. Such text provides potential for unobtrusive, inexpensive, and real-time measurement of psychological states (such as positive or negative affect) and aspects of subjective well-being (such as happiness and engagement). Social scientists have recently begun to use social media text in a variety of studies (Cohn et al., 2004; Kramer, 2010; Tausczik and Pennebaker, 2010; Kamvar and Harris, 2011; Dodds et al., 2011; Golder and Macy, 2011).

One of the most popular approaches to estimate psychological states is by using the *word count* method (Pennebaker et al., 2007), where one tracks the frequency of words that have been judged to be associated with a given state. Greater use of such words is taken to index the prevalence of the corresponding state. For example, the use of the word 'happy' is taken to index *positive emotion*, and 'angry' to index *negative emotion*. The most widely used tool to carry out such analysis, and the one we investigate in this paper, is Pennebaker's Linguistic Inquiry and Word Count, (*LIWC*) (Pennebaker et al., 2001; Pennebaker et al., 2007). *LIWC*, originally developed to analyze writing samples for emotion and control, has grown to include a variety of lexica for linguistic and psychosocial topics including positive and negative emotions, pronouns, money, work, and religion. The *word count* approach has high appeal to social scientists in need of a tool to approach social media, and although others have been used (see, for example (Gottschalk and Bechtel, 1998; Bollen et al., 2010), *LIWC*'s lexica are generally perceived as a "tried-and-tested" list of words (Miller, 2011).

Unfortunately, the *word count* approach has some drawbacks when used as indicators for psychological states. Words are the unit of measurement, but words can carry many different meanings depending on context. Consider the Facebook posts below containing instances of 'play', a word associated with positive emotion in *LIWC*.

*Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task*, pages 296–305, Atlanta, Georgia, June 13-14, 2013.

1. *so everyone should come to the **play** tomorrow...*

2. *Does anyone what type of file i need to convert youtube videos to **play** on PS3???*

3. *Time to go **play** with Chalk from the Easter Bunny!*

Out of the three instances, only (3) seems to communicate positive emotion. In (1), 'play' is used as a noun rather than the expected verb, while in (2), 'play' is a verb but it is used in a sense that is not directly associated with positive emotion. (1) and (2) demonstrate how *lexical ambiguities* (i.e. multiple parts-of-speech or word senses) can affect accuracy of words in a lexicon. Additionally, even when appearing as the expected part of speech and word sense, signal from a word may change due to its context, such as being within the scope of a negation as in (4), or describing something desired as in (5).

4. *...all work no **play** :-(*

5. *i sure wish i had about 50 hours a day to **play** cod*

Our goal is to characterize the errors of the widely used *word count* approach, and show that such lexica can be significantly improved by employing an ambiguity metric to refine such lexica. Rather than work on a new method of measuring psychological states, we work within the bounds of *word count* and ask how accurate it is and whether we can improve it without sacrificing its simplicity and scalability.

We attempt to reduce the erroneous signal of the *word count* approach while maintaining legitimate signal simply by refining the lexicon. In other words, we would like to move closer to the goal in Figure 1, by eliminating words that often carry erroneous signal such as 'play', and keeping words which often carry the sought-after signal, such as 'cheerful'. The difficulty in doing this is that we do not have the data to tell us which words are most likely to carry signal (even if we had such data we would like to develop a method that could be applied to any newly created lexica). Instead we leverage part-of-speech and word sense data to help us determine which words are lexically ambiguous.

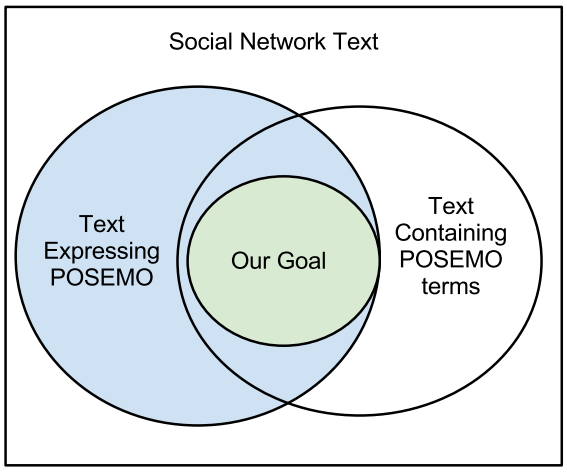

Figure 1: The relationship between text expressing positive emotion (*POSEMO*) and text containing *LIWC* terms for *POSEMO*.

Our approach of eliminating ambiguous words increases the precision at the expense of recall, a reasonable trade-off in social media where we are working with millions or even billions of word instances. Additionally, it is minimally-supervised, in that we do not require training data on human-state; instead we use existing hand-labeled corpora, such as SemCor (Miller et al., 1993), for word sense information. Not requiring training data also means our refinement is flexible; it can be applied to multiple domains and lexica, it makes few assumptions that might introduce problems of over-fitting, and it is parsimonious in that it merely improves an established approach.

This paper makes two primary contributions: (1) an analysis of the types of errors common for the *word count* approach (Section 3), and (2) a general method for refining psychosocial lexica based on the ambiguity of words (Section 4). Before describing these contributions, we discuss related work, making the case for using social media in social science and surveying some work in computational linguistics. We then evaluate both the original *LIWC* lexicon and our refinement of it against human judgments of expression of positive and negative emotions on hand-annotated Facebook posts, and show the benefit of lexicon refinement for estimating well-being over time for large aggregates of posts. Finally, we discuss the implications of our work and possible future directions.

2 Background

Compared to traditional approaches in the social sciences, large scale analysis of social media is cheap, near real-time, unobtrusive, and gives high coverage. We outline these advantages below.

Inexpensive Extracting information from sources such Facebook and Twitter is vastly cheaper than the more conventional polling done by companies such as Gallup – and by many social science researchers. Social media data does not require phone calls to be made or doors to be knocked on. For example, a representative survey asking 1,000 people by a leading polling company costs to the order of $10,000[1]. In contrast, once the software exists, social media data from tens of millions of users can be obtained and analyzed at a fraction of the cost.

Temporal Resolution Much of the attraction of social media stems from the fact that it captures a written live stream of collective thought. When Google relied on search queries to monitor health-seeking behavior to predict influenza epidemics, the reporting lag was a mere day, whereas traditional CDC surveillance systems take 1-2 weeks to publish their data (Ginsberg et al., 2009). Infrastructure based on social media and Internet use data allows reporting and analysis systems with little to no reporting lag. Additionally, traditional survey designs are typically only designed to assess psychological states at a given point in time.

Unobtrusive Estimation Traditional self-report survey approaches, even those implemented on the web, suffer from social desirability, priming, and other biases. For example, Kahneman et al. (Kahneman et al., 2006) found that the order in which questions are asked on questionnaires can determine how they are answered. By looking directly into the social worlds, many of these self-report biases can be avoided. The traces of human interactions in social media represent the goings-on in their original ecologies of meaning and signification. This approach diminishes the inferential distance between the context of the phenomena and the context of measurement – and thus decreases the room for systematic distortion of signal.

[1]Gallup, Personal correspondence.

2.1 The Word Count Approach

As previously noted, the *word count* approach is most often used by social scientists through the tool known as *Linguistic Inquiry and Word Count* or *LIWC* (Pennebaker et al., 2007). The LIWC2007 dictionary is composed of almost 4,500 words and word stems organized across one or more word categories, including 406 positive emotion words and 499 negative emotion words. When long form texts are analyzed with LIWC, the program simply returns the percentages of words belonging to the different analytical categories – the simplicity of this approach makes it popular with non-technical social scientists.

LIWC's positive and negative emotion lexica have recently begun to be used on "short form" writing in social media. For example, Golder and Macy (2011) used LIWC to study diurnal and seasonal variation in mood in a collection of 400 million Twitter messages. Kramer (2010) proposed the "Gross National Happiness" index and Kivran-Swaine and Naaman (2011) examined associations between user expressions of positive and negative emotions and the size and density of social networks. A comprehensive review can be found in Tausczik and Pennebaker (2010).

To our knowledge there is only one work which has evaluated *LIWC*'s accuracy over social media. Bantum and Owen (2009) evaluated *LIWC* on a set of posts to an Internet-based breast cancer support group. By annotating expression of emotion within this text, they were able to produce accuracy figures of *sensitivity* (much like *recall*) and *predictive validity* (*precision*). Sensitivity measured how often a word (in context) expressing positive or negative emotion was captured by *LIWC*. Predictive validity measured how often a word (in context) captured by *LIWC* as measuring positive or negative emotion was indeed expressing positive or negative emotion. While they found a recall of 0.88, the precision was only 0.31 – that is, only 31% of instances containing words indexed by *LIWC* actually conveyed the associated emotion. We contend that this is a major drawback for applying *LIWC* to social media, because while it is not important to catch every expression of emotion out of a million Tweets, it is important that when something is captured it is an accurate

estimate of the true state.

2.2 Related Work in Computational Linguistics

Researchers have been exploring the use of lexica that define the subjective orientation of words for tasks such as sentiment or subjectivity analysis. A common weakly-supervised approach starts with a small set of sentiment knowledge (seed words associated with a given sentiment) and expands the words into a large lexicon (Hatzivassiloglou and McKeown, 1997; Kamps and Marx, 2002; Kim and Hovy, 2004; Kanayama and Nasukawa, 2006; Baccianella et al., 2010). We take a different approach. Rather than expanding lexica, we start with a large set of words and refine the set. The refinement increases precision at the cost of recall, which is a reasonable exchange when we are looking at millions or even billions of word instances. Standard applications of sentiment analysis, such as annotating movie reviews, may not be as inclined to skip instances, since they want to make predictions for items which have very few reviews.

Another line of work in sentiment analysis has created lexicons using supervised learning. One of the first works to do so was by Pang and colleagues (2002), who used data including author ratings of reviews, such as IMDB movie reviews. The author ratings become training data for sentiment classification. Pang et al. showed that human-created lexicons did not perform as well as lexicons based on simple word statistics over the training data. Interestingly, they found that words like 'still' were most predictive of positive movie reviews, and that punctuation marks of '!' and '?' were strong signs of negative movie reviews. Unfortunately, training data for subjective well-being or happiness is not yet available, preventing the use of such supervised learning methods. Additionally, this work seeks to experiment within the bounds of what social scientists are in fact using (with publications in high-impact venues such as Science). We thus take a different approach, and automatically improve human created lexicons.

Wiebe and Cardie (2005) generalized the task of sentiment analysis to that of discovering subjectivity such as "opinions, emotions, sentiments, speculations, evaluations, etc.". More recently, Wilson et al. (2009) contended that the context may neutralize or change the polarity of the subjective orientation of a word. It is difficult to determine where concepts of happiness such as quality of relationships or degree of achievement in life fit in with subjectivity. Thus, we do not claim to be measuring subjectivity and instead we use the general term of 'psychological state', referring to "the way something [a person] is with respect to its main attributes" (Miller, 1993).

To the best of our knowledge, while part-of-speech tagging and word sense disambiguation are staple tasks in the computational linguistics community, the utility of a lexical ambiguity metric has yet to be explored.

POSEMO		NEGEMO	
term	**frequency**	**term**	**frequency**
like	774,663	hate	167,109
love	797,833	miss	158,274
good	571,687	bad	151,496
friend*	406,568	bore*	140,684
happy	384,797	shit*	114,923
LOL	370,613	hurt*	98,291
well*	284,002	craz*	94,518
great	263,113	lost	94,059
haha*	240,587	damn*	93,666
best	227,381	fuck	90,212
better	212,547	stupid*	85,587
fun	216,432	kill*	83,593
please*	174,597	hell	80,046
hope	170,998	fuckin*	79,959
thank	161,827	wrong*	70,714

Table 1: Most frequent *POSEMO* and *NEGEMO* terms in *LIWC* in the 12.7 million Facebook posts. "*" indicates a wildcard, so that "well*" matches "wellness".

3 Annotation and Analysis of Errors from the Word Count Method

One objective of our work is to document and describe how often different types of errors occur when using the *word count* approach on social media. To do this, we first judged a sample of 1,000 instances of *LIWC* terms occurring in Facebook posts to indicate whether they contribute signal towards the associated LIWC category (i.e. positive emotion). We then took instances that were deemed to carry erroneous signal and annotated them with a label for the

category	agreement	instances	base rate
POSEMO	0.742	500	.654
NEGEMO	0.746	500	.697
TOTAL	**0.744**	**1,000**	**.676**
random	*0.343*	-	-

Table 2: Inter-annotator agreement over 1,000 instances of *LIWC* terms in Facebook posts. Base rate is the average of how often an annotator answered true.

type of signal error. This section describes the process we used in generating these annotations and the results we found.

3.1 Annotation Process

Annotating social media instances of lexica terms provides insight into how well the *word count* approach works, and also yields a "ground truth" for evaluating our lexicon refinement methods. We randomly selected for labeling a sample of 1,000 status updates containing words from a given lexicon drawn from a collection of 12.7 million Facebook status updates provided by the Cambridge myPersonality project (Kosinski and Stillwell, 2012).

We used terms from the *LIWC* positive emotion (*POSEMO*) and negative emotion (*NEGEMO*) lexica, which are the same lexica used by the works of Kramer (2010), Kivran-Swaine and Naaman (2011), and Golder and Macy (2011). Table 1 lists the most frequent *POSEMO* and *NEGEMO* terms in our Facebook sample.

As mentioned above, we did two types of annotations. First, we judged whether each given instance of a word conveyed the correct associated type of emotion. The second task took a sample of instances judged to have incorrect signal and labeled them with a reason for the error; We refer to this as *signal error type*.

For the first task, we had three human judges independently evaluate the 1,000 status update instances as to whether they were indeed correct signal. The question the judges were told to answer was "Does the word contribute to the associated psychological-state (POSEMO or NEGEMO) within the sentence it appears?". In other words, "would the sentence convey less [positive emotion or negative emotion] without this word?". Subjective feedback from the judges indicated that it was often difficult to make a decision, so we used three judges per instance. In the case of conflict between judges, the "correct" label for validation of the refined lexicon was defined to be the majority vote. A sampling of Facebook statuses demonstrates a mixed picture of relevance for the unrefined *LIWC* dictionaries:

1. *has had a very **good** day* ('good' - *POSEMO*)
2. *is so very **bored**.* ('bore*' - *NEGEMO*)
3. ***damn**, that octopus is good, lol* ('damn' - *NEGEMO*)
4. *thank you for his **number*** ('numb*' - *NEGEMO*)
5. *I got pranked sooooo **bad*** ('bad' - *NEGEMO*)
6. *don't be **afraid** to fail* ('afraid' - *NEGEMO*)
7. *I wish I could . . . and we could all just be **happy*** ('happy' - *POSEMO*)

Some posts clearly use positive or negative lexicon words such as (1) and (2). Curse words can signify negative emotion or emphasize the opposite state as in (3), which is clearly emphasizing positive emotion here. Example (5) demonstrates the word sense issue we discussed previously. Words with wildcards that expand into other words with different meanings can be particularly problematic, as the expanded word can be far more frequent – and very different in meaning – from the original word. For example, 'numb*' matches 'number' in 4.

A different problem occurs when the context of the word changes its implication for the emotional state of the writer. This can either occur through negation such as in (6) where 'afraid' signals *NEGEMO*, but is negated with 'don't' or the signal can be changed indirectly through a variety of words indicating that the writer desires (and hence lacks) the state, as in (7) where someone is wishing to be 'happy'.

Table 2 shows the agreement between annotators calculated as $\frac{\sum_i agree(A_1^{(i)}, A_2^{(i)}, A_3^{(i)})}{1{,}000}$, where $agree(A_1, A_2, A_3)$ was 1 when all three annotations matched and 0 otherwise. Given the average positive base rate across annotators was 0.676 the chance that all three reviewers agree according to chance (random agreement) is calculated as

category	precision	instances
POSEMO	67.9%	500
NEGEMO	72.8%	500
both	**70.4%**	**1,000**

Table 4: Accuracy of *LIWC POSEMO* and *NEGEMO* lexica over Facebook posts.

$0.676^3 + (1-0.676)^3 = 0.343$, the probability of all three answering yes plus the probability of all three answering no.

For the second task, we selected 100 instances judged to be incorrect signal from the first task, and labeled them according to the best reason for the mistake. This task required more linguistic expertise and was performed by a single annotator. Labels and descriptions are given in Table 3, which breaks down the cases into lexical ambiguity, direct or indirect negation, and other reasons such as the stemming issue (stem plus wildcard expanding into words indicating a different (or no) emotional state).

3.2 Analysis of Errors

Before discussing the types of errors we found when using the *word count* approach, we examine *LIWC*'s overall accuracy on our dataset. Table 4 shows the precision broken down for both the positive emotion (*POSEMO*) and the negative emotion (*NEGEMO*) lexica. We see that the precision for *NEGEMO* is slightly higher than *POSEMO*, indicating the terms in that category may be more likely to indicate their associated state.

Although the overall accuracy seems decent, one should keep in mind our subjective judgement criteria were quite tolerant, allowing any amount of contribution of the corresponding signal to be considered accurate. For example, a salutation like "Happy New Year" was judged to be a correct use of "happy" to signal POSEMO, even though it clearly does not have as strong a signal as someone saying "I feel deliriously happy".

Frequencies of signal errors are given in Table 5. The most common signal error was wrong word sense, where the word did not signal emotional state and some other sense or definition of the word was intended (e.g. "u feel like ur **living** in a music video"; corresponding to the sense "to inhabit" rather than the intended sense, "to have life; be alive" (Miller, 1993)). Other common signal errors include strict negation where the word is canceled out by a clear negative quantifier (e.g. "Don't be **afraid** to fail") and wrong part of speech where the word is signaling a different part of speech than the emotion (e.g. "**well**, we cant afford to go to NYC"). There were also various other signal error types that include stem issues where the stem matched clearly unintended words, desiring statuses where the status is commenting on wanting the emotion instead of experiencing it and other less prevalent issues such as non-English language post, memes, or clear sarcasm.

category	label	frequency
Lexical Ambiguity	Wrong POS	15
	Wrong WS	38
Signal Negation	Strict Negation	16
	Desiring	6
Other	Stem Issue	5
	Other	24

Table 5: Frequency of the signal error types.

4 Method for Refining Lexica

The idea behind our refinement method is to remove words that are likely to carry erroneous signal about the underlying state or emotion of the person writing the tweet or Facebook post.[2] We do so in an indirect fashion, without actually using training data of which posts are, in fact indicative of positive or negative emotion. Instead, we focus on reducing errors that are due to lexical ambiguity. By removing words that are often used with multiple parts of speech or multiple senses, we can tilt the balance toward precision at some cost in recall (losing some signal from the ambiguous words). This makes the *word count* approach more suitable for use in the massive corpora afforded by social media.

4.1 Lexical Ambiguity

We address lexical ambiguity at the levels of both part of speech (POS) and word sense. As a metric of inverse-ambiguity, we determine the probability that a random instance is the most frequent sense (*mfs*) of the most frequent part of speech (*mfp*) of

[2]Refinement tool is available at `wwbp.org`.

category	label	description	examples
Lexical Ambiguity	Wrong POS	Not a valid signal because it is the wrong POS	*so everyone should come to the **play** tomorrow...*
	Wrong WS	Not a valid signal because it is the wrong word sense (includes metaphorical senses)	*Does anyone what type of file i need to convert youtube videos to **play** on PS3???*
Signal Negation	Strict Negation	Within the scope of a negation, where there is a clear negative quantifier	*...all work **no play** :-(*
	Desiring	Within the scope of a desire / wishing for something	*i sure wish i had about 50 hours a day to **play** cod*
Other	Stem Issue	Clearly not intended to be matched with the given stem	**numb*** for *NEGEMO* matching *number*
	Other	Any other issue or difficult to classify	

Table 3: Signal error types.

the word, denoted TSP (for *top sense probability*). Given a word w, we consider all parts of speech of w ($POS(w)$) and all senses for the most frequent part of speech ($senses(\textit{mfp}(w))$):

$$p_{\textit{mfp}}(w) = \frac{\max_{[w_{pos} \in POS(w)]} f_p(w_{pos})}{\sum_{w_{pos} \in POS(w)} f_p(w_{pos})}$$

$$p_{\textit{mfs}}(w) = \frac{\max_{[w_{sense} \in senses(\textit{mfp}(w))]} f_s(w_{sense})}{\sum_{w_{sense} \in senses(\textit{mfp}(w))} f_s(w_{sense})}$$

$$TSP(w) = (p_{\textit{mfp}}(w) * p_{\textit{mfs}}(w))^2 \quad (1)$$

Here, f_p and f_s represent the frequencies of a certain part-of-speech and a certain sense of a word, respectively. This is the squared-probability that an instance of w is the *top sense* – the most-frequent part-of-speech and the most-frequency sense of that part-of-speech. The probability is squared because both the word in the lexicon and the word occurring in context should be the *top sense* (two independent probabilities: given an instance of a word in a corpus, and another instance of the word in the lexicon, what is the probability that both are the top POS and sense?). Frequency data is provided for parts-of-speech from the Google N-Grams 2.0 (Lin et al., 2010) and for word senses from SemCor (Miller et al., 1993). This aspect of the refinement is inspired by the most frequent sense heuristic for word sense disambiguation (McCarthy et al., 2004; Yarowsky, 1993), in which the sense of a word is chosen without regard to the context, but rather is simply based on the frequencies of senses in corpora. In our case, we restrict ourselves this way in order for the application of the lexicon to remain unchanged.

For some words, we were unable to find sense frequency data. We decided to keep such terms, on the assumption that a lack in available frequency information implies that the word is not very ambiguous. Many of these terms include Web speak such as 'haha' or 'lol', which we believe can carry a strong signal for positive and negative emotion.

Lastly, since TSP is only a metric for the inverse ambiguity of a word, we must apply a threshold to determine which words to keep. We denote this threshold as θ, and the description of the refined lexicon for a category, cat, is below.

$$lex_\theta(cat) = \{w | w \in cat \wedge TSP(w) > \theta\}$$

4.2 Handling Stems

Some lexica, such as the *LIWC* dictionary, include word stems that are intended to match multiple forms of a word. Stems are marked by the suffix '*'. *LIWC* describes the application of stems as follows "the asterisk, then, denotes the acceptance of all letters, hyphens, or numbers following its ap-

lex	cat	prec	size
full	POSEMO	67.9%	500
	NEGEMO	72.8%	500
	both	**70.4%**	**1,000**
lex$_{0.10}$	POSEMO	70.9%	392
	NEGEMO	71.6%	423
	both	**71.3%**	**815**
lex$_{0.50}$	POSEMO	75.7%	239
	NEGEMO	78.9%	232
	both	**77.3%**	**471**
lex$_{0.90}$	POSEMO	72.5%	109
	NEGEMO	78.1%	128
	both	**75.5%**	**237**

Table 6: Precision (**prec**) and instance subset size (**size**) of refinements to the *LIWC POSEMO* and *NEGEMO* lexica with various θ thresholds $(0.10, 0.50, 0.90)$

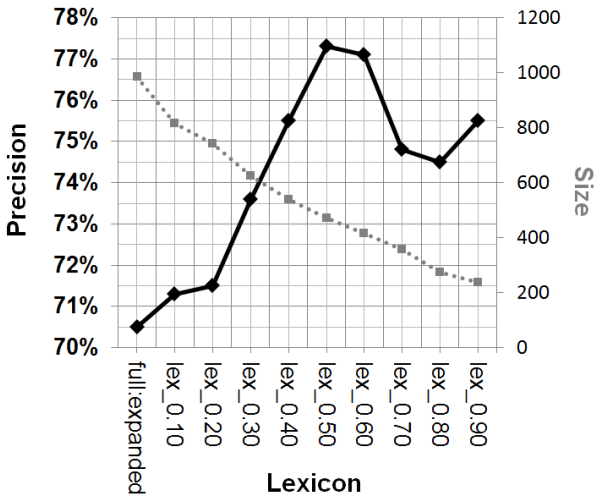

Figure 2: The relationship between precision and size when increasing the TSP threshold (θ).

pearance."[3] This presents a problem because, while the creators of such lexica obviously intended stems to match multiple forms of a word, stems also often match completely different words, such as 'numb*' matching 'number' or 'won*' matching 'won't'.

We identified how often unintended matches happen in Section 3. Finding that the stemming issues were not the biggest problem, here, we just describe how they fit into our lexical ambiguity metric, rather than describe a technique to rid the lexicon of stemming problems. One approach might be to determine how ambiguous a stem is – i.e. determine how many words, parts-of-speech, and senses a stem could be expanded into, but this ignores the fact that the dictionary creators obviously intended the stem to match multiple words. Instead, we expand stems into all words that they match and replace them into the lexica.

We base our expansion on the actual terms used in social media. We find all words matching stems among 1 million randomly selected Twitter messages posted over a 6-month period (August 2009 - February 2010), and restrict to those occurring at least 20 times. Then, each word stem in the lexicon is replaced with the expanded set of matching words.

5 Evaluation

We evaluate our refinement by comparing against human judgements of the emotion conveyed by words in individual posts. In the case of human judgements, we find that the subset of human-annotated instances matching the refined lexica are more accurate than the complete set.

In section 3 we discussed the method we used to judge instances of *LIWC POSEMO* and *NEGEMO* words as to whether they contributed the associated affect. Each of the 1,000 instances in our evaluation corpus were judged three times such that the majority was taken as truth. In order to validate our refined lexica, we find the accuracy (precision) of the subset of instances which contain the refined lexica terms.

Table 6 shows the change in precision when using the refined lexica. **size** represents the number of instances from the full evaluation corpus matching words in the refined lexica. One can see that initially precision increase as the **size** becomes smaller. This is more clearly seen in Figure 2. As discussed in the method section, our goal with the refinement is improving precision, making lexica more suitable to applications over massive social media where one can more readily afford to skip instances (i.e. smaller size) in order to achieve more accuracy. Still, removing more ambiguous words does

[3] "How it works": http://www.liwc.net/howliwcworks.php

not guarantee improved precision at capturing the intended psychological state; it is possible that that all senses of an ambiguous word do in fact carry intended signal or that the intended sense a low ambiguity word is not the most frequent.

Our maximum precision occurs with a threshold of 0.50, where things somewhat level-out. This represents approximately a 23% reduction in error, and verifies that we can increase precision through the automatic lexicon refinement based on lexical ambiguity.

6 Conclusions

Social scientists and other researchers are starting to measure psychological states such as happiness through text in Facebook and Twitter. We have shown that the widely used *word count* method, where one simply counts occurrences of positive or negative words, can often produce noisy and inaccurate estimates of expressions of psychological states. We characterized and measured the frequency of different types of errors that occur using this approach, and found that when counting words without considering context, it is lexical ambiguities (unintended parts-of-speech or word senses) which cause the most errors. We proposed a method for refining lexica by removing those words most likely to be ambiguous, and showed that we can significantly reduce error as measured by human judgements.

Acknowledgments

Support for this research was provided by the Robert Wood Johnson Foundation's Pioneer Portfolio, through a grant to Martin Seligman, "Exploring Concepts of Positive Health". We thank the reviewers for their constructive and insightful comments.

References

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may. European Language Resources Association (ELRA).

Erin O.C. Bantum and J.E. Owen. 2009. Evaluating the validity of computerized content analysis programs for identification of emotional expression in cancer narratives. *Psychological assessment*, 21(1):79.

Johan Bollen, Huina Mao, and Xiao-Jun Zeng. 2010. Twitter mood predicts the stock market. *Computer and Information Science*, 1010:1–8.

Michael A. Cohn, M.R. Mehl, and J.W. Pennebaker. 2004. Linguistic markers of psychological change surrounding september 11, 2001. *Psychological Science*, 15(10):687.

Peter Sheridan Dodds, Kameron Decker Harris, Isabel M Kloumann, Catherine A Bliss, and Christopher M Danforth. 2011. Temporal patterns of happiness and information in a global social network: Hedonometrics and twitter. *Diversity*, page 26.

Jeremy Ginsberg, M.H. Mohebbi, R.S. Patel, L. Brammer, M.S. Smolinski, L. Brilliant, et al. 2009. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–4.

Scott A. Golder and M.W. Macy. 2011. Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science*, 333(6051):1878–1881.

Louis A. Gottschalk and RJ Bechtel. 1998. Psychiatric content analysis and diagnosis (pcad2000).

Vasileios Hatzivassiloglou and Kathleen R. McKeown. 1997. Predicting the semantic orientation of adjectives. In *Annual Meeting of the Association for Computational Linguistics*, pages 174–181.

Daniel Kahneman, A.B. Krueger, D. Schkade, N. Schwarz, and A.A. Stone. 2006. Would you be happier if you were richer? a focusing illusion. *Science*, 312(5782):1908.

Jaap Kamps and Maarten Marx. 2002. Words with attitude. In *1st International WordNet Conference*, pages 332–341, Mysore, India.

Sepandar D. Kamvar and J. Harris. 2011. We feel fine and searching the emotional web. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 117–126. ACM.

Hiroshi Kanayama and Tetsuya Nasukawa. 2006. Fully automatic lexicon expansion for domain-oriented sentiment analysis. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, pages 355–363, Stroudsburg, PA, USA. Association for Computational Linguistics.

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

Funda Kivran-Swaine and M. Naaman. 2011. Network properties and social sharing of emotions in social awareness streams. In *Proceedings of the ACM 2011*

conference on Computer supported cooperative work, pages 379–382. ACM.

Michal. Kosinski and David J. Stillwell. 2012. mypersonality research wiki. mypersonality project. *http://www.mypersonality.org/wiki/*.

Adam D.I. Kramer. 2010. An unobtrusive behavioral model of gross national happiness. In *Proceedings of the 28th international conference on Human factors in computing systems*, pages 287–290. ACM.

Dekang Lin, Kenneth Church, Heng Ji, Satoshi Sekine, David Yarowsky, Shane Bergsma, Kailash Patil, Emily Pitler, Rachel Lathbury, Vikram Rao, Kapil Dalwani, and Sushant Narsale. 2010. New tools for web-scale n-grams. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may. European Language Resources Association (ELRA).

Diana McCarthy, Rob Koeling, Julie Weeds, and John Carroll. 2004. Finding predominant word senses in untagged text. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04), Main Volume*, pages 279–286, Barcelona, Spain, July.

George A. Miller, Claudia Leacock, Randee Tengi, and Ross T Bunker. 1993. A semantic concordance. In *Proceedings of the ARPA Workshop on Human Language Technology*, pages 303–308. Morgan Kaufman.

George A. Miller. 1993. Five papers on wordnet. *Technical Report, Princeton University*.

Greg Miller. 2011. Social scientists wade into the tweet stream. *Science*, 333(6051):1814–1815.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 79–86.

James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: Liwc 2001. *Word Journal Of The International Linguistic Association*.

James W. Pennebaker, C.K. Chung, M. Ireland, A. Gonzales, and R.J. Booth. 2007. The development and psychometric properties of liwc2007. *Austin, TX, LIWC. Net*.

Yla R. Tausczik and J.W. Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24.

Janyce Wiebe and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. language resources and evaluation. In *Language Resources and Evaluation*.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2009. Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. *Computational Linguistics*, 35:399–433, September.

David Yarowsky. 1993. One sense per collocation. In *Proceedings of the workshop on Human Language Technology*, HLT '93, pages 266–271, Stroudsburg, PA, USA. Association for Computational Linguistics.