Empath: A Framework for Evaluating Entity-Level Sentiment Analysis

Charles B. Ward, Yejin Choi, Steven Skiena
Computer Science Dept.
Stony Brook University
Stony Brook, NY, USA
{charles, ychoi, skiena}@cs.sunysb.edu

Eduardo C. Xavier
Institute of Computing
University of Campinas (UNICAMP)
Campinas, SP, Brazil
ecx@ic.unicamp.br

***Abstract*—Sentiment analysis is the fundamental component in text-driven monitoring or forecasting systems, where the general sentiment towards real-world entities (e.g., people, products, organizations) are analyzed based on the sentiment signals embedded in a myriad of web text available today. Building such systems involves several practically important problems, from data cleansing (e.g., boilerplate removal, web-spam detection), and sentiment analysis at individual *mention-level* (e.g., phrase-, sentence-, document-level) to the aggregation of sentiment for each *entity-level* (e.g., person, company) analysis. Most previous research in sentiment analysis however, has focused only on individual *mention-level* analysis, and there has been relatively less work that copes with other practically important problems for enabling a large-scale sentiment monitoring system.**

In this paper, we propose *Empath*, a new framework for evaluating *entity-level* sentiment analysis. *Empath* leverages objective measurements of entities in various domains such as people, companies, countries, movies, and sports, to facilitate *entity-level* sentiment analysis and tracking. We demonstrate the utility of *Empath* for the evaluation of a large-scale sentiment system by applying it to various lexicons using *Lydia*, our own large scale text-analytics tool, over a corpus consisting of more than a terabyte of newspaper data. We expect that *Empath* will encourage research that encompasses end-to-end pipelines to enable a large-scale text-driven monitoring and forecasting systems.

I. Introduction

Sentiment analysis is the fundamental component in text-driven monitoring or forecasting systems, where general sentiment towards real-world entities (e.g., people, products, organizations) are analyzed based on the sentiment signals embedded in a myriad of web text available today. One popular application of sentiment analysis has been to extract consumer opinions from product reviews (e.g., [12], [20], [34]). More complex and yet exciting applications of sentiment analysis include predicting stock price movement, public opinion of politicians, and the spread on football games from a large corpus of news articles [18], [19], [11]. Building such text-driven forecasting systems requires dealing with several practically important problems, from data cleansing (e.g., boilerplate removal, web-spam detection), to sentiment analysis at individual *mention-level* (e.g., phrase-, sentence-, document-level) to cross-document or even cross-corpus aggregation of mention-level sentiment for each real-world entity. In this paper, we call this aggregated sentiment for each real-world entity *entity-level* sentiment analysis.

Most previous research in sentiment analysis, however, has focused only on individual mention-level analysis where significant progress has been made for subjectivity detection as well as polarity classification. Subjectivity detection concerns the existence of opinion or sentiment in text, while polarity classification concerns the overall positivity or negativity of text (e.g., whether a review of a product is favorable). Indeed, there has been a great deal of progress made for polarity classification (e.g., , [6], [23], [29]) and subjectivity identification (e.g., [15], [31]) at all level of textual mentions: phrase-, sentence-, and document-level. However, there has been relatively less work that cope with other practically important problems for enabling a large-scale sentiment monitoring system.

In this paper, we tackle *real-world* entity-level sentiment analysis directly. That is, given a set of text documents for a specific range of dates, we aim to extract named entities and determine their sentiment polarity for the given time period. Given news data, countries with low infant mortality should have higher sentiment than countries with high infant mortality, good movies should have higher sentiment than bad movies, and Nobel Laureates should have higher sentiment than criminals. Although mention-level sentiment analysis is a critical component here, other aspects of the system, such as data cleansing and aggregation, also affect the overall system performance for entity-level sentiment analysis.

We argue that in order to boost research activities on global *entity-level* sentiment analysis, we need a commonly-shared evaluation framework that directly measures the quality of the system at the entity-level, rather than just mention-level. Although it is not completely unreasonable to imagine that a sentiment analysis algorithm that works substantially better at *mention-level* would also work better at *entity-level*, no previous work has reported such correlation between mention-level performance and entity-level performance. In fact, it might be the case that a sentiment analysis algorithm that works slightly better at the *mention-level* (e.g., by 2%) does

978-1-4577-1591-4/11/$26.00

Data Standard File	Generated queries file	User answers file
Infant Mortality, 2006, ranking, decreasing		
Iceland, 20060101, 20061231, 2.9	Iceland, 20060101, 20061231	Iceland, 38621, 24052, -16218
Singapore, 20060101, 20061231, 3.0	Singapore, 20060101, 20061231	Singapore, 1241, 776, -266
...	...	...
Afghanistan, 20060101, 20061231, 157	Afghanistan, 20060101, 20061231	Afghanistan, 1138607, 783560, -1411157
Sierra Leone, 20060101, 20061231, 160.3	Sierra Leone, 20060101, 20061231	Sierra Leone, 43775, 33665, -65750

TABLE I
(LEFT) AN EXAMPLE OF DATA STANDARD FILE. (CENTER) THE RESULTING SET OF QUERIES GIVEN TO THE SENTIMENT SYSTEM UNDER EVALUATION. (RIGHT) SENTIMENT DATA RETURNED FROM THE SYSTEM.

not make any difference at *entity-level*. After all, in order to perform well at entity-level, the algorithm does not need to understand every single mention or occurrence of sentiment correctly. Rather, it is important to recognize the representative and prominent sentiment correctly. Below we summarize the potential weaknesses of using only mention-level evaluation strategies with respect to text-driven sentiment forecasting systems:

- **Incomplete System Evaluation** – Evaluation based only on a mention-level gold standard is not able to provide insight into end-to-end system performance. For example, mention-level evaluation cannot quantify end-to-end system improvement with respect to (a) spam/duplicate elimination, (b) data source reputation evaluation, and (c) robustness to different document streams (e.g. news articles vs. Twitter vs. translated text).
- **Annotation Cost** – Although some of the mention-level annotation can be derived automatically, e.g., movie reviews [23] and product reviews, there are many other domains for which such automatic acquisition of gold standard is not trivial, such as general news articles and blogs. Manual annotation is known to be very expensive and time-consuming to produce. As a result, studies involving the hand-annotation of more than a few thousand articles are rare.
- **Alternate Text Streams** – The best document-oriented gold standards focus tightly on specific domains (e.g. movie reviews), and are almost certainly mono-lingual. Such standards have limited meaning beyond the single language / domain they are derived for.

Addressing the above concerns, we present *Empath*, a evaluation framework which provides a complementary sentiment evaluation strategy, global *entity-level* evaluation in particular, based on correlations with independent statistical data that reflect underlying sentiment in representative news text. For evaluation purposes, we invert the task of text-driven forecasting: instead of predicting real events by using a sentiment system, we seek to evaluate the quality of a sentiment system using the a-priori sentiment of real entities over specified time periods.

For example, consider the example *Empath* standard in Table I, showing the infant mortality rate for various countries in 2006. Even in contexts not related to infant mortality, we would believe that over a large news text corpus, the infant mortality rate for these countries should correlate with their entity-level sentiment.

This approach avoids the cost and complication of creating human-annotated text corpora. More importantly, we can evaluate the end-to-end performance of a sentiment system by measuring the final sentiment signal at the aggregate entity level. Improvements to sentiment assignment, spam detection, source reliability measures, etc., will be reflected in the score provided by *Empath*.

The primary contributions of this work are:

- **End-to-end evaluation** – We develop a supplementary approach for evaluating sentiment systems which does not rely on manually annotated document corpora. By collecting a broad class of signals and world events which have relatively unambiguous sentiment interpretation, we can assess the end-to-end system performance of a sentiment system.
- **Evaluation of the** $Lydia$ **sentiment system** – We demonstrate the value of the *Empath* evaluation environment, performing experiments using our own sentiment system, *Lydia* [2], [21], [10], [1]. over a terabyte-scale corpus of over 100 million U.S. daily newspaper articles from 2004-2010. *Lydia* is a powerful system, and its sentiment analysis has been used for successful text-driven forecasting projects [33], [11], [32]. We show that our sentiment evaluation environment allows us to study the impact of design decisions for sentiment analysis systems. Specifically, we demonstrate how *Empath* can be used to evaluate two aspects of the *Lydia* sentiment system: (1) the lexicon of sentiment words used by the *Lydia* sentiment system, and (2) the manner in which the *Lydia* system links sentiment words to entities.

A. Related Work

Most of the work on evaluating a sentiment analysis system is based on human annotated polarity of texts such as [17], [28], or using other manually assigned open sources such as Epinions, Amazon etc. Manually annotated testbeds, however, involve significant effort. For this reason, studies involving hand-annotation of more than a few thousand articles are rare. Open sources such as Amazon and Epinions prove valuable, although not perfect, as studies show that users are biased to give positive scores [13], [24].

Some previous work on sentiment analysis avoids the use of manually-annotated text. Ghose et. al [9] uses econometrics to derive polarity and strength of opinions of eBay seller reviews by tracing the change in economic variables. Koppel and Shtrimberg [16] used a similar approach to annotate the Mutex

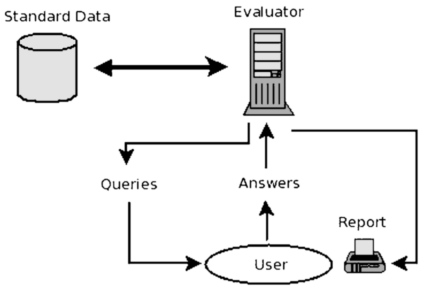

Fig. 1. Architecture of the *Empath* System.

Significant Developments corpus stories as being negative or positive based in the stock price change of the previous day. Devitt and Ahmad [7] proposes the use of external measures such as economic variables and human behavior (using psychology methods to detect it) while humans are reading a text, and then use these measures to evaluate polarity of the texts. Our approach in evaluating a sentiment analysis system is related to these previous approaches, since we also use objective data to evaluate polarity of entities.

Several methods have been proposed to generate sentiment lexicons, typically by applying a machine-learning technology to an appropriate linguistic resource (e.g. WordNet or an annotated corpus). One common approach involves the construction of lexicons of sentiment bearing words [10], [25], [34], using seed sets of relevant sentiment words and expanding using WordNet [22] synonym and antonym relations. Several sentiment lexicons have been published and are available for comparative analysis, including SentiWordNet [25], ANEW [3] and DAL [27]. More recently optimization ILP methods have been used for the purpose of generating sentiment lexicons [5].

II. The *Empath* System

Here we describe the architecture of *Empath*, the objective standards presently incorporated in the system, and evaluation methods it employs. The overall architecture of *Empath* is simple, (see Fig. 1), consisting of:

- **Entity Standards** – Objective measures of the sentiment of entities in various domains, including businesses, companies, movie stars, Nobel Laureates, and criminals. These standards can be scores, such as reputation rankings of businesses, or an objective binary polarity classification (criminals vs. Nobel Laureates). *Empath* currently has standards for more than $5,000$ entities.
- **System Evaluator** – Given sentiment standards, the sentiment system evaluator generates sentiment queries which the sentiment system being evaluated must answer. These answers are then processed and scored by the evaluator.

As described earlier, Table I gives an example of the type of standards, queries, and answers used by *Empath*.

A. Standards

Empath employs sentiment standards for various entities with substantial variety in both entity type and domain. These standards fall into three distinct types.

- **Class Standards**: Entities grouped into the categories good or bad. For example, criminals are bad and should have low sentiment polarity scores, while Nobel Laureates are good and should have high sentiment polarity scores.
- **Rankings Standards**: Lists of entities ranked by a standard variable, such as countries by GDP or movies by IMDB rating.
- **Time-Series Standards**: Entity time-series of a standard variable, such as stock price over time.

The five primary categories of standards in *Empath* we report on are:

- **Business:** We use the yearly Reputation Institute survey of domestic and international companies reputation variable as a ranking standard and the Fortune Magazine fifty most admired companies by year as a class standard.
- **Stock Market:** We use monthly price data for shares of each company on the Dow Jones Index as a time-series standard, computed as both the average share price for the month and as an overall change for the month.
- **Movies:** We take the top fifty grossing movies of each year as ranking tests, as ranked by gross and IMDB rating.
- **People:** We consider lists of Nobel laureates, Oscar winners, Time Magazine heroes and icons, the FBI most wanted list,and executed criminals as class based tests.
- **Countries:** We rank countries by annual UN statistics of GDP per capita, Human Development Index index, and infant mortality rate.

We will discuss the details of how each of these test types are evaluated in the following section.

B. Evaluation Model Rationale

Given the set of standard files, the evaluator generates a series of "queries" which are to be answered by the sentiment analysis system under evaluation. These queries consist of an entity name and a date range, and the system is expected to return the sentiment data for the entity over the given date range. Sentiment polarity is usually measured as a unit-less quantity; in $Lydia$, as a value between -1 and 1, with 0 indicating neutral sentiment. It is an appealing general assumption to require only a polarity value for each entity.

Unfortunately, when the corpus size from which the scores are being computed is sufficiently large, we effectively measure only the precision and not the recall of the sentiment system. That is, given an enormous volume of relevant news text to analyze, even a very low-recall system will perform well. The tradeoff between precision and recall for the sentiment system can therefore not be properly evaluated.

To illustrate this issue, imagine two sentiment systems Q and $Q^{'}$ which compute sentiment polarity scores by scoring individual sentences. Suppose sentiment system $Q^{'}$ is identical to Q but that it randomly ignores 90% of its input data.

Suppose Q is asked to evaluate the polarity of *Enron* over a period of one year. For such a query there may be many thousand relevant sentences. Because a small fraction of the data is still a large dataset, $Q^{'}$ will, by the law of large numbers, yield a very similar sentiment polarity score for *Enron* for this period, with high probability. Nonetheless, Q is clearly a better sentiment system than $Q^{'}$, and will capture the negative sentiment associated with *Enron* after having seen much less data. It is clear that our evaluation method must capture this.

Thus, what we need to evaluate a sentiment analysis system is a probability distribution of polarity scores for each entity, given a random subset of the corpus of a fixed size X (choosing X small enough to elucidate the performance differences of systems, e.g., 20 sentences containing the entity in question). This distribution would specify the probability that the entity will be assigned each particular polarity score, given a random size X subset of the corpus. We would then use such a probability distribution to score the probability of the system giving correct answers for the appropriate query type, e.g., in the case of class based queries, the probability of the system placing the entity in the correct sentiment class.

C. Evaluation Model

It is unreasonable, however, to demand such complex output from the sentiment system under evaluation. Therefore, we seek a proxy which sentiment systems should generally be able to provide. Instead of merely a polarity score, the system under evaluation must supply an ordered triple specifying: (a) the number of occurrences of the entity in the input corpus, (b) the number of positive scorings and (c) negative scorings of that entity over these occurrences. See Table I (right) for an example input for a small set of entities.

Note that this model is still very general, as the quanta of sentiment analysis can be words, sentences, or articles. In the case of $Lydia$, the number of occurrences corresponds to sentences, while positive/negative scorings denote the linkage of individual sentiment words to the entity. Alternatively, article granularity could be substituted without fundamentally altering the model. Of course, scores from systems which apply different meanings to these inputs cannot be fairly compared.

The evaluator now analytically computes probability distributions of positive and negative occurrences, based on the assumption that positive and negative scorings are (a) equally weighted, and (b) independently and randomly distributed among occurrences. Specifically, we assume that positive and negative occurrences are each Poisson distributed, with rates derived from the number of positive and negative scorings per entity occurrence. Thus, for our fixed X occurrences, we can compute a matrix of probabilities for each pair of possible positive and negative scorings.

We will use the following definition of sentiment polarity score (sp) based on the number of positive (pos) and negative (neg) occurrences:

$$sp = \frac{pos - neg}{pos + neg}$$

This polarity score will range from -1 to 1, with 0 representing an equal number of positive and negative references. The three types of tests in *Empath* are:

a) Class Tests: Class tests are the simplest case. Recall that in a class based test we wish to ascribe the correct sentiment polarity to an entity categorized as either positive or negative. One important note is that although the $Lydia$ lexicon is balanced in that positive and negative sentiment words occur with entities at very nearly equal rates, this is not true for several of our tested lexicons. We define the positive / negative threshold t as the average entity score in our dataset on that lexicon.

Let P_{cor} (P_{inc}) be the probability that the sentiment polarity score is not within an arbitrary ϵ of t and the assigned polarity of the entity is correct (incorrect). Let P_{neu} be the probability that the sentiment polarity score is within ϵ of the neutral threshold t. Therefore, $P_{cor} + P_{inc} + P_{neu} = 1$. Then, we define precision and recall for a single entity as:

$$Precision = \frac{P_{cor}}{P_{cor} + P_{incor}} \qquad Recall = P_{cor}$$

This way of defining precision and recall is in line with the probabilistic notion of precision and recall in the information retrieval (IR) context. In IR, probabilistic precision is defined as the probability that a retrieved document is relevant, while probabilistic recall is the probability that a relevant document will be retrieved. In our case, retrieved documents are non-neutral sentiment assignments of entities, while relevant documents are the set of all correct entity assignments. Thus, precision is the probability that a non-neutral sentiment assignment is a correct sentiment assignment, and recall is the probability that a correct sentiment assignment is assigned.

To aggregate the results across all entities, we simply average these precision and recall scores.

b) Ranking Tests: Recall that in ranking tests, we have a set of entities rank ordered by some standard variable, e.g., countries ordered by infant mortality. For each pair of entities, let P_{cor} be the probability that the sentiment polarity score correctly orders this pair of entities, e.g., that a country with higher infant mortality has a lower sentiment score than a country with lower infant mortality. Similarly, let P_{inc} be the probability that they are incorrectly ordered and P_{neu} be the probability that their polarity scores are the same.

We now define precision and recall for each pair in the same way as for class tests. We use a weighted average to aggregate precision and recall results across all pairs. The weight of each pair is based on the difference in rank between two entities. That is, it is more important that very different entities be distinguished correctly than relatively similar entities.

c) Time Series Tests: In time series tests we are given a set of standard values for a single entity over various time periods. For example, changes in a company's stock price should be reflected in the company's sentiment. We treat each time series test as a single ranking test. The time periods are rank ordered according to the standard value, and precision and recall are computed as in a ranking test.

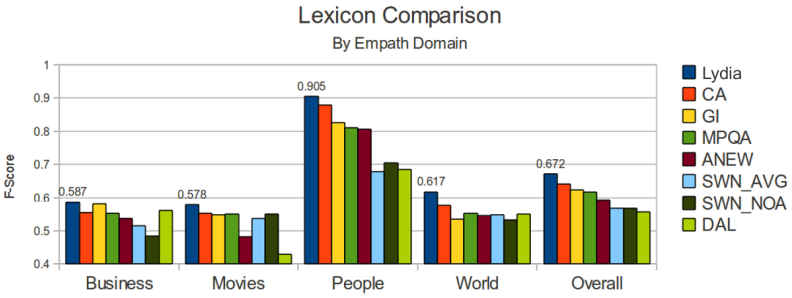

Fig. 2. F-score of various lexicons, by domain of test standard.

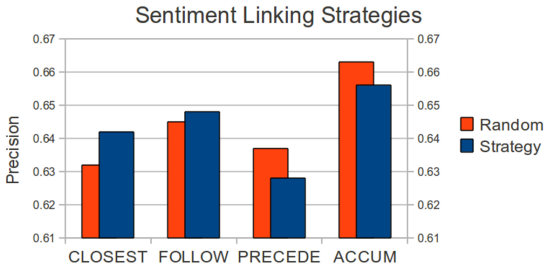

Fig. 3. Linkage subsets – Various linkage subsets compared to equivalently sized random linkage selections. The four linkage strategies are described in section III-B

III. *Lydia* Experiments with *Empath*

The *Lydia* sentiment analysis system [2], [21], [10], [1], applies a given lexicon L to sentences, linking sentiment words from L to entities in sentences. The default lexicon used by *Lydia* was derived by its creators using WordNet, and has several domain-specific sub-lexicons. *Lydia* has been applied to several very large corpora, providing insights into areas as diverse as sociology [26], political science [4], and text-driven forecasting [33], [11], [32]

The default linking method used by *Lydia* applies every sentiment word in the sentence to every entity in the sentence, with a limited set of negation ("not bad") and emphasis ("very bad") criteria. Both the lexicon and the linking method used by *Lydia* are completely modular, allowing us to perform experiments using various lexicons and linking methods. In this section we will describe the results of testing different sentiment lexicons and sentiment word linking strategies using *Empath*.

A. Sentiment Lexicon Comparison

We evaluate a number of published sentiment lexicons:

- *Lydia* **default** – The default *Lydia* lexicon was created by expanding seed sets of words from various domains using WordNet [22].
- **SentiWordNet (SWN)** – SentiWordNet [8] uses several different learning algorithms to compute objectiveness and polarity of word synsets in WordNet. We consider this lexicon using two methods of converting it to an unweighted lexicon: (1) we average the number of positive and negative classifications across all word senses with the same part of speech ("SWN_AVG"), and (2) we remove all potentially ambiguous words ("SWN_NOA").
- **Affective Norms for English Words (ANEW)** – ANEW [3] is a dictionary of words with human annotated ratings in three dimensions. We use the "pleasant" to "unpleasant" dimension of this dictionary to create a lexicon of positive and negative words.
- **Dictionary of Affect in Language (DAL)** – The Dictionary of Affect in Language (DAL) [27] is also human annotated dictionary; like ANEW, we use DAL's "pleasant" to "unpleasant" dimension.
- **Opinion Finder (MPQA)** – We use the polarity values from the subjectivity lexicon used in OpinionFinder [30].
- **General Inquirer (GI)** – [14] provides word lists tagged according to different dimensions; for our purposes here, we use the polarity dimension.

In addition to these published lexicons, we also created lexicons which attempt to combine the information in multiple lexicons. As there is substantial disagreement between lexicons, we took several variations on resolving these disagreements. The best performing lexicon of this group (CA), was created by combining the three best performing lexicons (*Lydia*, GI, and MPQA); a word which appears in multiple lexicons with different polarities was assigned the polarity of the majority, and removed if tied. Variations which used more dictionaries and/or sought more complete agreement between dictionaries performed similarly, but slightly more poorly.

After running each lexicon through *Lydia*, we obtain the *Empath* results shown in Figure 2. We are relieved to find that the default *Lydia* lexicon performs better in every domain overall, surpassing even the consensus lexicon. Other interesting effects can be observed: e.g., both human annotated dictionaries of "pleasantness" do quite poorly in the movie domain.

B. Sentiment Word Linking

In its default configuration, *Lydia* assigns every sentiment word which appears in a sentence to every entity in that sentence. Although this linking method may seem simplistic, as we will show, it is a non-trivial problem to achieve sufficient improvement in the precision of the sentiment words linked to an entity to justify the resulting loss of quantity of linked words. We evaluate several linkage heuristics, striving for more accurate assignment:

- **Closest** – Each entity is linked to the single sentiment word which is closest to it in the sentence (CLOSEST).
- **Following** – Each entity is linked to every sentiment word which follows it, up to the next entity (FOLLOW).
- **Preceding** – Each entity is linked to every sentiment word which precedes it, up to the previous entity (PRECEDE).
- **Accumulate** – Each entity is linked to every sentiment word in the sentence which appears before it (ACCUM).

We compare each of these strategies to a baseline of selecting an equivalent percentage of random linkages (15.6% for Closest, 24.4% for Following, 18.4% for Preceding, and 44.5% for Accumulate). The results of these four strategies are shown in Figure 3.

None of these linking strategies are better, in absolute performance, than the baseline: simply linking every sentiment word to every entity. However, the Closest and Following strategies perform substantially better than an equivalent random subset. In contrast, Preceding and Accumulate do worse than their equivalent samples.

IV. Conclusion

Our experiments show that certain lexicons and sentiment words are more informative than others, and *Empath* can help us identify these. Unfortunately, it is also clear that we cannot choose only the most informative data points, as the result is a system whose individual signals are more precise, but whose output signals are less so. Any data source that carries signal ultimately should not be ignored, but instead weighted according to its value.

For future work, we are working to evaluate more sentiment signals, and to re-engineer the $Lydia$ system to incorporate a broader flexible class of weightings. Weighted lexicons are a part of this, but we must also weigh data sources by reliability, sentiment linkages by their position, and sentences by their complexity.

References

[1] M. Bautin, L. Vijayarenu, and S. Skiena. International sentiment analysis for news and blogs, 2008.

[2] M. Bautin, C. Ward, A. Patil, and S. Skiena. Access: News and blog analysis for the social sciences. In *19th Int. World Wide Web Conference (WWW 2010)*, Raleigh NC, 2010.

[3] M. M. Bradley and J. P. Lang. Affective norms for english words (anew): Instruction manual and affective ratings. Technical report, The Center for Research in Psychophysiology, University of Florida, 1999.

[4] S. Carey, M. Lebo, and S. Skiena. Leading, following or informing: online news and its impact on british public attitudes. Nuffield College Political Science Seminar Series, University of Oxford, 11 November 2008.

[5] Y. Choi and C. Cardie. Adapting a polarity lexicon using integer linear programming for domain-specific sentiment classification. In *Conference on Empirical Methods in Natural Language Processing*, pages 590–598, 2009.

[6] K. Dave, S. Lawrence, and D. M. Pennock. Mining the peanut gallery: opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th international conference on World Wide Web*, WWW '03, pages 519–528, New York, NY, USA, 2003. ACM.

[7] A. Devitt and K. Ahmad. Sentiment analysis and the use of extrinsic datasets in evaluation. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, 2008.

[8] A. Esuli and F. Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. In *Language Resources and Evaluation (LREC)*, 2006.

[9] A. Ghose, P. G. Ipeirotis, and A. Sundararajan. Opinion mining using econometrics: A case study on reputation systems. In *ACL*, pages 416–423, 2007.

[10] N. Godbole, M. Srinivasaiah, and S. Skiena. Large-scale sentiment analysis for news and blogs. In *Proc. First Int. Conf. on Weblogs and Social Media*, pages 219–222, 2007.

[11] Y. Hong and S. Skiena. The wisdom of bookies? sentiment analysis versus. the nfl point spread. In *International AAAI Conference on Weblogs and Social Media*, 2010.

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

[13] N. Hu, P. A. Pavlou, and J. Zhang. Can online reviews reveal a product's true quality?: empirical findings and analytical modeling of online word-of-mouth communication. In *ACM Conference on Electronic Commerce*, pages 324–330, 2006.

[14] G. Inquirer. http://www.wjh.harvard.edu/ inquirer/, 2002.

[15] S.-M. Kim and E. Hovy. Automatic detection of opinion bearing words and sentences. In *IJCNLP*, pages 61–66, New York, 2005. Springer.

[16] M. Koppel and I. Shtrimberg. Good news or bad news? let the market decide. *Computing Attitude and Affect in Text: Theory and Applications*, 20(1):297–301, 2006.

[17] L.-W. Ku, Y.-S. Lo, and H.-H. Chen. Test collection selection and gold standard generation for a multiply-annotated opinion corpus. In *ACL*, pages 89–92, 2007.

[18] V. Lavrenko, M. Schmill, D. Lawrie, P. Ogilvie, D. Jensen, and J. Allan. Mining of concurrent text and time series. In *In proceedings of the 6 th ACM SIGKDD Int'l Conference on Knowledge Discovery and Data Mining Workshop on Text Mining*, pages 37–44, 2000.

[19] K. Lerman, A. Gilder, M. Dredze, and F. Pereira. Reading the markets: forecasting public opinion of political candidates by news analysis. In *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1*, COLING '08, pages 473–480, 2008.

[20] B. Liu, M. Hu, and J. Cheng. Opinion observer: analyzing and comparing opinions on the web. In *Proceedings of the 14th international conference on World Wide Web*, WWW '05, pages 342–351, New York, NY, USA, 2005. ACM.

[21] L. Lloyd, D. Kechagias, and S. Skiena. Lydia: A system for large-scale news analysis. In *String Processing and Information Retrieval (SPIRE 2005)*, pages 161–166, 2005.

[22] G. A. Miller. WordNet: a lexical database for English. *Commun. ACM*, 38(11):39–41, 1995.

[23] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *EMNLP '02: Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, pages 79–86, Morristown, NJ, USA, 2002. Association for Computational Linguistics.

[24] P. Resnick and R. Zeckhauser. The value of reputation on ebay: A controlled experiment. *Experimental Economics*, 9(2):79–101, 2006.

[25] F. Sebastiani and A. Esuli. Determining term subjectivity and term orientation for opinion mining. In *In Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 193–200, 2006.

[26] A. van de Rijt, E. Shor, C. Ward, and S. Skiena. Only fifteen minutes? the social immobility of fame in english-language newspapers. Under Review, 2011.

[27] C. M. Whissell. *The dictionary of affect in language*, pages 113–131. 1989.

[28] J. Wiebe, T. Wilson, and C. Cardie. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2/3):164–210, 2005.

[29] T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 347–354, 2005.

[30] T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *HLT-EMNLP-2005*, 2005.

[31] H. Yu and V. Hatzivassiloglou. Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 129–136, Morristown, NJ, USA, 2003. Association for Computational Linguistics.

[32] W. Zhang and S. Skiena. Improving movie gross prediction through news analysis. In *Web Intelligence*, pages 301–304, 2009.

[33] W. Zhang and S. Skiena. Trading strategies to exploit blog and news sentiment. In *International AAAI Conference on Weblogs and Social Media*, 2010.

[34] L. Zhuang, F. Jing, and X.-Y. Zhu. Movie review mining and summarization. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, CIKM '06, pages 43–50, New York, NY, USA, 2006. ACM.