Coupling Niche Browsers and Affect Analysis for an Opinion Mining Application

Gregory Grefenstette, Yan Qu, James G. Shanahan, David A. Evans
Clairvoyance Corporation
5001 Baum Bd, Suite 700
Pittsburgh, PA, 15213-1854, USA
{grefen,yqu,jimi,dae}@clairvoyancecorp.com

Abstract

Newspapers generally attempt to present the news objectively. But textual affect analysis shows that many words carry positive or negative emotional charge. In this article, we show that coupling niche browsing technology and affect analysis technology allows us to create a new application that measures the slant in opinion given to public figures in the popular press.

Introduction

Although major newspapers generally strive for objectiveness in reporting, it is nearly impossible to use words which do not carry some emotional content when describing an event or a person. By coupling a niche browser, GoogleNews, which extracts temporally ranked news items from the Web, with Affect Analysis technology, we can find underlying nuance and slant in otherwise objective news-wire text. After a description of Niche Browsers and Affect Analysis, we present our system which couples both in order to provide an entity-directed opinion miner. We test our miner on text from sites which we know to be favorable or unfavorable to certain entities, by examining its results over left and right-wing political figures in conservative and liberal newspaper.

Niche Browsers

Niche browsers produce full-text indexes of documents found on the web, such as AllTheWeb and Google, but, rather than indexing all the pages that they find, niche browsers first classify pages and then only index pages corresponding to a specific class of papers. Two popular niche browsers are the ResearchIndex (http://citeseer.nj.nec.com/cs), well known in scientific circles, which piggy-backs on general search engines (Lawrence et al, 1999); and Google News, which limits its browsing to 4,500 online news sources[1].

Niche browsers differ from online interfaces to classified databases, e.g., Medline (http://www.ncbi.nlm.nih.gov/PubMed), which provide indexed access to proprietary databases, because although information in both cases is restricted to a certain class of information, niche browsers (i) must deal with a wide variety of input document formats, (ii) gather their input documents from the changing Web rather than from a locally controlled database, and (iii) have no control over document field definition.

Since niche browsers do not control the format of the documents they must analyze, most depend on entity extraction technology (Appelt, 1999; Cunnigham, 1999). For example, FlipDog (www.flipdog.com) "crawls the World Wide Web and links to job openings found on employer Web

[1] http://news.google.com/help/about_news_search.html

sites" by recognizing entities such as geographical locations, job titles, salary offerings, educational requirements, and company and people names. Steele (2001) describes some other niche browsers such as the home page search engine HPSearch, MySimon which does price analysis, MoreOver which analyzes new sources, as well as describing differences between specialized engines and general web browsers. PROGENIE (Duboué *et al.*, 2003) proposes to extract biographical information from free-text sources.

Affect Analysis

Affect analysis is a natural language processing technique for recognizing the emotive aspect of text. The same textual content can be presented with different emotional slants. For example, one description of actors in a civil war may be described as *freedom-fighters* whereas another describing the same events may use *terrorists*, the first term having a positive connotation and the second, a negative connotation.

In the 1960s, Stone and Lasswell began building lexicons in which words were labeled with affect. In the Lasswell Value Dictionary (Lasswell and Namenwirth, 1969), the word *admire*, for example, was tagged with a positive value along the affect dimension *RESPECT*. This dictionary marked words with binary values along eight basic value dimensions (WEALTH, POWER, RECTITUDE, RESPECT, ENLIGHTENMENT, SKILL, AFFECTION, and WELLBEING). Stone's work on the General Inquirer dictionary (Stone, *et al*. 1965) has continued to this day[2]. The dictionary now (early 2004) contains 1,915 words marked as generally positive and 2,291 words as negative. A wide variety of other affect classes are used to label entries, e.g., *Active, Passive, Strong, Weak, Pleasure, Pain, Feeling (other than pleasure or pain), Arousal, Virtue, Vice, Overstated, Understated*. In addition to these labels, an open-ended of set of semantic labels have been defined, e.g., *Human, Animate, …, Region, Route,…, Object, Vehicle,…, Fetch, Stay,…* In these dictionaries, all labels are binary. For example, in the current General Inquirer, the word *admire* possesses among its labels *Positive* and *Pleasure*. Words either possess an attribute or not; there is no question of degree.

In addition to these manually constructed lexicons that include affect attributes, work has begin on automatically acquiring affect information. Hatzivassiloglou & McKeown (1997) demonstrated that, given a set of emotively charged adjectives, positively oriented adjectives tended to be conjoined to positively oriented adjectives, and negative adjectives to negative ones, in expressions such as "good and honest" or "bad and deceitful." In their experiments, they took a number of frequently occurring adjectives that they decided had some type of orientation and then used statistics on whether two adjectives appeared together in a corpus in the pattern *X and Y* to decide if they had the same orientation. They created graphs in which the nodes were the adjectives and links between nodes showed that they appeared in the patterns; they then divided this graph into two groups with the minimum number of links between the groups. The bigger class of words was considered as having negative polarity (since there are more negative words than positive words in English). They achieved 92% accuracy over a test set of 236 adjectives that they classified as positive or negative.

In other research, Weibe (2000) used a seed set of "subjective" adjectives and a thesaurus generation method (Hindle, 1990) to find more subjective adjectives. Turney & Littman (2003) have shown how it was possible to automatically discover positively and negatively charged words, given fourteen seed words, and using statistics of association from the WWW. Their set of positive charged seed words was {*good, nice, excellent, positive, fortunate, correct, superior*} and their set of negatively charged seed words was {*bad, nasty, poor, negative, unfortunate, wrong, inferior*}. They found that positive words tend to associate more often with the positive words than with the negative word, using a form of pointwise mutual information (Church & Hanks, 1989) and page statistics on word appearance on

[2] At http://www.wjh.harvard.edu/~inquirer/inqdict.txt, one can find an online version of this dictionary.

Altavista. For example, if one wished to decide if a word like *fortuitous* was positive or negative, then one could send requests such as "fortuitous NEAR good" and "fortuitous NEAR bad," using the NEAR operator and the advanced search facility of Altavista.com. Using this method, they achieved 98.2% accuracy with the 334 most frequently found adjectives in the Hatzivassiloglou and McKeown test set. They also showed that it should be possible to reliably extend a list of positively and negatively marked words such as that found in the General Inquirer lexicon (Stone, 1966).

In addition to merely tagging affect-laden terms as positive or negative, one can also position affect words along more discrimination axes. For example, one might consider that not only adjectives but nouns carry affect, and that words like *adore, affection, ardor, caress, embrace, …* are also associated more or less stronger and with more or less intensity along an axis that one could label as *love*. At www.humanityquest.com, one finds a list of more than 500 different human values which could serve as such axes.

In the late 1990s, we began development of a lexicon of affect words by hand (Subasic and Huettner, 2000a, 2000b, 2001; Huettner and Subasic 2000). Entries in our lexicon consist of five fields: (i) a lemmatized word form, (ii) a simplified part of speech [adjective, noun, verb, adverb], (iii) an affect class, (iv) a weight for the centrality of that word in that class, and (v) a weight for the intensity of the word in that class. The centrality of a word is a hand-assigned value between 0.0 and 1.0 that is intended to capture the relatedness of the word to the affect class. The intensity value attempts to capture the emotional strength of the word. For example in the sample entries given below, we see that the adjective *gleeful* has been assigned to two affect classes (*happiness* and *excitement*) and that it has been deemed more related to the class *happiness*, with a centrality of 0.7 than it is to the class *excitement* where the lexicon creators only gave it a 0.3 intensity:

"gleeful" adj happiness 0.7 0.6
"gleeful" adj excitement 0.3 0.6

In both entries, the word *gleeful* was deemed to have an intensity of 0.6 (out of a maximum intensity of 1). The combination of intensities and centralities made it possible to develop multidimensional weightings of affect in texts (Subasic and Huettner 2000a, 2000b, 2001). Our lexicon contains 2258 words (corresponding to 4926 surface forms) that are classed into 83 affect classes.

In the work described below, we use a simplified version of this affect lexicon. Similar to the work on positively and negatively charged words in the research mentioned above, we have a version of our affect lexicon in which each class (such as *happiness*) is labeled as positive or negative. We use this simple binary tagging of affect classes in what follows. This simplified version looks like the following. The first column contains the affect word, the second contains one of the classes the word has been assigned to, and the third contains a positive/negative sign associated with that class.

admonish warning -
admonishment warning -
admonition warning -
adorable attraction +
adoration love +
adoration superiority +
adore love +
adulterer immorality -
adultery immorality -
advantage advantage +
advantage superiority +

Coupling to Create an Entity Directed Opinion Miner

In the following sections, we describe our application which couples affect analysis with a niche browser, specifically the GoogleNews browser, in order to create a time-bounded entity-directed opinion miner.

Our system functions as follows:

- The end-user specifies the entity about whom the current public opinion (as voiced in the press) is to be mined, as well as the time period involved.
- Our system sends a request to the Google News browser and fetches up to 1000 references to news articles concerning this entity during the specified period.
- Each article is fetched, and the text around the specified entity is extracted (using a KWIC Keyword-in-Context (Heaps, 1978) program). We used 120 characters before and after the entity as a window.
- The extracted windows are sorted and duplicates removed (to eliminate duplicate articles portions).
- The windows are collated, and all affect words (in any morphological variant) from our lexicon are identified. Affect classes are associated with each affect word using the lexicon.
- A score for the entity is produced by dividing the number of instances of a positive affect class by the number of instances of a negative affect class. If there are more positive than negative references, then, the score will be greater than 1; if there are more negative references, it will be less than one.

Example

In August 2003, we applied the system by extracting opinion around "Qusay Hussein," following his death. The system was run using the following command:

./getnews "Qusay Hussein" "Qusay"

The first string after the command *getnews* was sent to Google News to retrieve articles containing this string. By default the 1000 most recent articles mentioning "Qusay Hussein" were retrieved, though it is possible to also specify a range of dates. The second string "Qusay" was used for the KWIC window extraction in the retrieved articles. This distinction between the two strings is useful because the first string allows for more accurate retrieval while the second string can be a string corresponding to the shorter form by which the entity is referred throughout the article.

For example, in articles extracted using "Qusay Hussein", the KWIC program extracted windows such as the following, centered on the string "Qusay":

> *… detaining scores of people. Saddam's feared sons Uday and Qusay were buried on Saturday on the outskirts of Tikrit ….*
> *… most people don't know about and gave details on the final countdown of the end of Uday's and Qusay's reign of terror…*

These windows were sorted and duplicates eliminated. In the remaining segments, all affect words from our lexicon were identified, e.g. *detain, feared, terror,* and assigned to their affect classes via lookup in the affect lexicon. No disambiguation was performed to decide which affect class to assign if more than one could be assigned, but words which had ambiguous aspect, i.e., which belonged to both positively and negatively charged classes, were removed from consideration. The following classes, preceded by their frequency of evocation, were found in these windows (a plus or minus after the class name shows whether the class is considered positive or negative):

467 violence -
400 sadness -
399 fear -
361 conflict -
302 death -
237 excitement +
228 force -
206 horror -
189 slyness -
184 avoidance -

The actual affect words (both positive and negative) that are found in these words are (in decreasing frequency) : *buried, dead, battle, deaths, hideout, feared, death, dead, dictator, vowed, attacked, died, burial, attack, betrayed, defeated, informant, intelligence, funeral, capture, corpses, fight, captured, order, fugitive, fall, combat, hope, force, fierce, powerful, gun, secret, grave, good, double, betrayal, promised, assault, approved, torture, help, disguise, grenades, friend, service, hoped, clear, clash, wanted, prevented, hammered, boasted, bad, imposed, hated, fighting, continue, concern, avenge, anger, guarded, grave, discovered, pressure, ordered, increase, handcuffed, flood, …*

In the final step, the score was assigned to "Qusay" by taking the counting number of instances of positively charged affect classes (1536) and the number of instances of negatively charged classes (3736) evoked in the retrieved text around "Qusay" and taking their ratio which yields 1536/3736 = 0.41. If this ratio is greater than 1, then we found more instances of positive affect class words in the news articles; if it is less than 1 then there were more negative class words present. For "Qusay Hussein" in this period, there were more than twice as many negatively charged word in the surrounding text than positive words.

Evaluation

Although it might seem historically clear that "Qusay" should be associated with negative affect words in that period of time, the above example is merely anecdotal. We performed the following experiment to validate whether the scoring process, simple as it is, corresponded to real slant. For this evaluation, we manually extracted two collections of text which, we felt, would present "George Bush" in a positive light for one collection and negative light in the other collection. For the positive text, we extracted pages mentioning "George Bush" from http://whitehouse.gov, the official site for the White House. For the negative text, we chose an anti-government, anti-Bush site "From the Wilderness" found at http://www.copvcia.com. From both sites, we extracted 100 pages (excluding pages referring to Mrs. Bush). Each collection was submitted to the process described above.

Running the window extraction and affect scoring steps over these two collections, we found the following scores for "Bush:"

WhiteHouse.gov	2.60
Copvcia.com	0.93

These scores show that the White House site uses positively charged words in the windows of text around "Bush" whereas the anti-government site uses more negatively charged affect words (since the ration is less than one) in such windows. The White House site delivers the following affect words in the

neighborhood of "Bush": *decoration, not_fear*[3]*, help, capture, attack, right, laugh, guard, force, combat, improve, good, dog, disaster, disarm, destruction, aid, wise, struggle, service, proud, prosperous, order, love, increased, important, hope, honored, grave, friend, foreign, fighting, expect, ensure, effective, dent, damage, challenge, assistance, appeal, agree, successes, success, strong, spirit, secure, safe, responsive, responsible, real, prosecute, prevent, prevail, plain, peace, not_terror, not_responsible, not_just, intelligence, integration, infected, hostages, honored, heart, health, great, grave, genocide, game, funerals, frustrated, frank, fire, enhance, educational, disasters, depart, demonstration, deliver, defend, creation, created, continue, confront, compassion, brave, assist,…*

On the anti-government site copviacia.com, we find the following words associated (in decreasing frequency) with Bush: *secret, force,, intelligence, attack, order, trust, invasion, power, impeachment, foreign, court, clear, union, resolution, fall, criminal, blood, reason, high, concerned, supreme, real, failed, death, allies, ordered, hard, coup, concern, appeal, advised, warning, trial, powerful, important, illegal, dangerous, want, not_doubt, justice, flight, fight, favored, experience, elites, doomed, dead, critical, crimes, crashed, crash, convicted, climax, benefit, assault, warned, reason, production, hit, health, great, sins, sense, scandal, promised, pressure, oversight, open, not_surprise, not_attack, lust, laugh,…*

The simple ratio scoring of positives over negatives in these two cases corresponds to our intuitions of the relative tones evoked when reading the list of words above.

As another test towards validating this scoring method, we considered two news sources and compared the treatment that they gave to two public personalities, president *George Bush* and *Howard Dean*, in the last two weeks of December 2003, during the beginning of the Democratic primaries when it seemed that Howard Dean was going to be the likely candidate facing George Bush in the upcoming 2004 presidential elections. We drew stories concerning these figures from two online sources: a conservative newspaper, the Washington Times, and a closer-to-the-center mainstream newspaper, the Washington Post. We applied our affect scoring system to news stories from each source. The results of our method give the scores in the table below. Though *George Bush* gets a slightly positive slant in both the *Post* and the *Times*, the conservative paper presents more liberal, Democrat *Howard Dean* in a predominantly negative fashion:

scores	Washington Post	Washington Times
Howard Dean	1.40	1.12
George Bush	1.15	1.40

Table 1: The scores give the ratio of positive words to negative words in a window of 240 characters around each candidate's name. A score of 1.0 means that there are as many positive affect words found as negative affect words. The *Washington Post* is considered as a newspaper in the center of the political spectrum, while the *Washington Times* has more conservative, Republican leanings. The scores show that both Dean (Democrat) and Bush (Republican) had more positive words found in their contexts in both journals, but there was higher density of positive words around Bush in the *Times* and a higher density around Dean in the *Post*.

If we narrow the window of text used around each name, we get the scoring behavior shown in the table below, in which we see that, in the *Washington Times*, with decreasing window sizes, a progressively greater proportion of positively charged words is associated with the name *George Bush*[4].

[3] *Not_fear* shows that we have implemented a limited scoping of negation. Any occurrence of *no, not, without* causes affect words found in the following five words to be rewritten with a *not_* prefix. The positive or negative orientation of the prefixed word is the opposite of the original word. *Not_fear*, thus, has a positive orientation.

[4] In order to verify whether readers are really left with a positive or negative impression, and determine what window size correlates best with human readers' impressions, one would need a controlled experiment involving human subjects. We leave this as our future work.

Window size	150 chars	100 chars	50 chars
Howard Dean *Wash Post*	1.12	1.16	1.19
George Bush *Wash Post*	0.93	1.13	1.17
Howard Dean *Wash Times*	1.09	1.15	1.17
George Bush *Wash Times*	1.45	1.72	2.50

Table 2: The scores give the ratio of positive words to negative words in windows of decreasing sizes around each candidate's name. A score of 1.0 means that there are as many positive affect words found as negative affect words. As the window narrows around the name *Bush* in the conservative paper the *Washington Times*, one finds 5 positively charged words for every two negatively charged words.

Press Slant over Time

In August, 2003, Arnold Schwarzenegger was running for governor of California, against Gray Davis who was being recalled by the Californian electorate. During that period we applied our entity-directed opinion miner to both candidates. At the time, our scorer gave the following scores:

Arnold Schwarzenegger 2.17
Gray Davis 1.14

This result is not surprising for *Gray Davis* since one of the reasons for his recall were a number of bad economic problems facing California. But the high scores for *Schwarzenegger* show that the text nearest his name in those same articles was much more positive than negative in the time leading up to the election that he was to win. After the election, this effect drops off. When we apply the same technique to articles concerning the, now, Governor Schwarzenegger, in December 2003, we find now

Arnold Schwarzenegger 1.32

We see that the words associated with *Schwarzenegger* are still positive but not as much so as during the campaign.

Related Work

There has been much related work in opinion mining. A number of researchers have taken sets of hand-scored reviews, such as those found on Amazon, and have tried to predict the rating that would be given to the review, using only the text of the review (Schein, et al., 2002; Kushal et al. 2003). Our work differs from these in that they use as training examples entire user reviews, which have been written specifically to give opinion one way or the other about the book, movie or product being reviewed, whereas our system deals with newspaper articles which can cover many subjects and persons, and which are usually written as to not express an opinion. Some researchers (Wiebe, 1994; Wiebe *et al*., 2003) have worked on trying to identify whose opinions are the statements found in an article, i.e., whether some statement being cited as the opinion of the article author or being cited as the opinion of some named person in the article. Their research identifies opinion by recognizing rhetorical structure, a more complicated, and as yet unresolved, process than the simple lexical matching we perform here.

Future Work

A number of parameters have been fixed in our system; e.g., we have taken the most recent 1000 news articles available, and used a window of 120 characters before and 120 characters after the entity. It would be interesting to perform human subject tests, presenting them with KWIC snippets and asking them judge the tone. This data could then be used to examine the optimal window needed. We are also working on improving our affect lexicon (Grefenstette *et al.*, 2004), work that could improve the sensitivity of the method.

Conclusion

We have presented a new application for discovering whether a person is being presented in a positive or negative light in public press during a period of time. This application has been made possible by coupling affect analysis technology and niche browsing technology. We use the niche browser, GoogleNews, to classify web sites into news stories and to select pages corresponding to a given period of time. Then we apply the affect analysis and a simple entity extractor to score the affect carrying words found around the entity in question. In this initial exploratory work, we have shown that our scoring methods seem to provide intuitively correct results for text that we believe to be for or against a certain personality. We have also shown that our scoring shows that conservative newspapers provide more positive views for conservative public figures. In contrast to existing recommender systems or automatic rating systems, our application produces scores without training and provides scores for entities in non-opinion-based text (i.e., newswire). Our system, coupling these two technologies, provides a rough method for uncovering nuance and slant in otherwise objective text.

Bibliographical References

Appelt, D. (1999). An Introduction to Information Extraction. *Artificial Intelligence Communications*, 12(3), 161−172.

Cunningham, H. (1999). Information Extraction: A user Guide. *Research Report CS-9907*, Department of Computer Science, University of Sheffield.

Duboué, P.A, McKeown, K., and Hatzivassiloglou, V. (2003) PROGENIE: Biographical Descriptions for Intelligence Analysis.In *Proceedings of the NSF/NIJ Symposium on Intelligence and Security Informatics*: 343−345

Grefenstette, G., Qu, Y., Evans, D.A., and Shanahan, J.G. (2004) Validating the Coverage of Lexical Resources for Affect Analysis and Automatically Classifying New Words along Semantic Axes. In *AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications (AAAI-EAAT 2004)* Stanford. March 22-24.

Hatzivassiloglou, V. and McKeown, K. (1997*).* Predicting the semantic orientation of adjectives. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL-EACL 97)*, 174−181.

Heaps, P. (1978) *Information Retrieval, Computational and Theoretical Aspects*, Academic Press, New York.

Huettner, A. and Subasic, P. (2000). Fuzzy Typing for Document Management, *ACL 2000* Software Demonstration, Hong Kong, October.

Kushal, D., Lawrence, S., and Pennock, D. (2003). Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews. In *Proceedings of WWW2003,* Budapest, Hungary, May 20-24, 519−528.

Lawrence, S., Bollacker, S.K. & Giles, C.L.. (1999) Indexing and Retrieval of Scientific Literature. In *Proceedings of the English International Conference on Information and Knowledge Managements*, 139−146.

Schein A. I., Popescul, A., and Ungar, L. H. (2002). Methods and Metrics for Cold-Start Recommendations. In *Proceedings of the XXV Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Tampere, Finland.

Steele, R. (2001) Techniques for Specialized Search Engines, In *Proceedings of Internet Computing '01*, Las Vegas, June 25-28.

Stone, P. J., Dunphy, D. C., Smith, M. S., and Ogilvie, D. M. (1966). *The General Inquirer: A Computer Approach to Content Analysis*. MIT Press, Cambridge, MA.

Subasic, P. and Huettner, A. (2000a) Affect Analysis of Text Using Fuzzy Semantic Typing, In *Proceedings of FUZZ-IEEE 2000*, San Antonio, May.

Subasic, P. and Huettner, A. (2000b) Calculus of Fuzzy Semantic Typing for Qualitative Analysis of Text, In *Proceedings of ACM KDD 2000, Workshop on Text Mining*, Boston.

Subasic, P. and Huettner, A. (2001) Affect Analysis of Text Using Fuzzy Semantic Typing, *IEEE Transactions on Fuzzy Systems*, Special Issue.

Turney, P.D. and Littman, M.L. (2003), Measuring praise and criticism: Inference of semantic orientation from association, *ACM Transactions on Information Systems (TOIS)*, 21 (4), 315−346

Wiebe, J.M. (1994), "Tracking Point of View in Narrative," Comp. Ling. 20: 233−287.

Wiebe, J.M., Breck, E., Buckley, C., Cardie, C., Davis, P., Fraser, B. Litman, D., Pierce, D., Riloff, D., Wilson, T., Day, D., and Maybury, M. (2003). Recognizing and Organizing Opinions Expressed in the World Press. In *Papers from the AAAI Spring Symposium on New Directions in Question Answering (AAAI tech report SS-03-07)*