Automated Sentiment Analysis

Eugene Butler

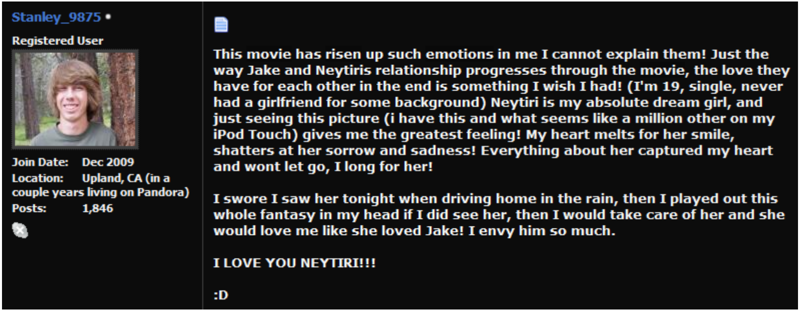
December 4, 2010

Introduction: Motivation

Traditional means of judging public opinion, such as customer satisfaction surveys, are time-consuming, expensive, and fraught with logistical difficulties.

- Difficult to reach a sufficiently large survey population.
- Mailed surveys are of limited effectiveness.
- Manual analysis of opinion takes too long.

Introduction: Solution



Stanley_9875

Registered User

Join Date: Dec 2009
Location: Upland, CA (in a couple years living on Pandora)
Posts: 1,846

This movie has risen up such emotions in me I cannot explain them! Just the way Jake and Neytiris relationship progresses through the movie, the love they have for each other in the end is something I wish I had! (I'm 19, single, never had a girlfriend for some background) Neytiri is my absolute dream girl, and just seeing this picture (i have this and what seems like a million other on my iPod Touch) gives me the greatest feeling! My heart melts for her smile, shatters at her sorrow and sadness! Everything about her captured my heart and wont let go, I long for her!

I swore I saw her tonight when driving home in the rain, then I played out this whole fantasy in my head if I did see her, then I would take care of her and she would love me like she loved Jake! I envy him so much.

I LOVE YOU NEYTIRI!!!

:D

- The internet is an increasingly popular venue for freely given opinion information.
- Online customer feedback influences the decisions of other customers.

Introduction: Modes of Automated Sentiment Analysis

Automated Sentiment Analysis refers to the computerized determination of the attitude of the author.

The purpose of this presentation is to examine the efficacies of four modern means of sentiment analysis:

1. Semantic Orientation from Pointwise Mutual Information (SO-PMI)
2. Semantic Orientation from Latent Semantic Analysis (SO-LSA)
3. Adjective Conjunction Measurement
4. Natural Language Processing Combined Method

Semantic Association: Known Words

Semantic association is predicated on the idea that “a word is characterized by the company it keeps.” In order to be able to find the semantic orientation of a new word, its positivity or negativity, one must first have a list of “known-positive” or “known-negative” words against which it can be compared.

- Good, Nice, Excellent, Positive, Fortunate, Correct, Superior,
- Bad, Nasty, Poor, Negative, Unfortunate, Wrong, Inferior.

Semantic Association: Corpora

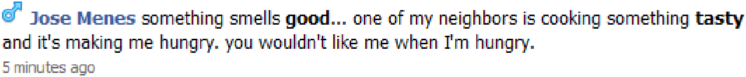
In order to compare our “known” words with neighbors, one must also provide a *corpus*:

- A newspaper. (30,000 words)
- Touchstone Applied Science Associates (TASA) set of short English documents. (10 million words)
- AltaVista Advanced Search engine English language page index. (100 billion words)

Larger corpora provide more information, at a price.

Semantic Association: Examples

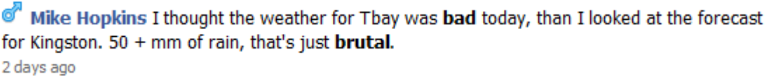
- Tasty



Jose Menes something smells **good...** one of my neighbors is cooking something **tasty** and it's making me hungry. you wouldn't like me when I'm hungry.

5 minutes ago

- Brutal

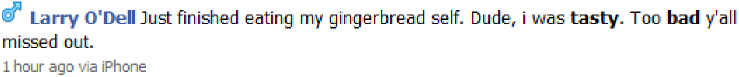


Mike Hopkins I thought the weather for Tbay was **bad** today, than I looked at the forecast for Kingston. 50 + mm of rain, that's just **brutal**.

2 days ago

Semantic Association: Examples

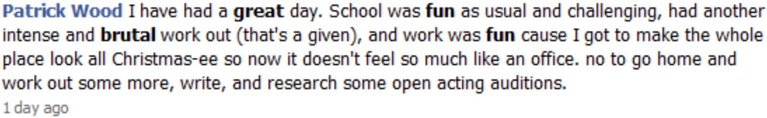
- Tasty



Larry O'Dell Just finished eating my gingerbread self. Dude, i was **tasty**. Too **bad** y'all missed out.

1 hour ago via iPhone

- Brutal



Patrick Wood I have had a **great** day. School was **fun** as usual and challenging, had another intense and **brutal** work out (that's a given), and work was **fun** cause I got to make the whole place look all Christmas-ee so now it doesn't feel so much like an office. no to go home and work out some more, write, and research some open acting auditions.

1 day ago

Large corpora helpful in establishing general word association.

Semantic Orientation from Association: Calculation

$$\text{SO-A}(w) = \sum_{p \in P} A(w, p) - \sum_{n \in N} A(w, n)$$

where w is the word we're interested in, p is a positive word, n is a negative word, and P and N are the sets of known positive and negative words.

The sum of these, for all w, equals the semantic orientation of our document. The absolute value represents our confidence in this orientation.

Semantic Orientation from Pointwise Mutual Information (SO-PMI)

In statistics, Pointwise Mutual Information is a way to measure the association of two outcomes by considering their coincidence (or lack thereof).

$$\mathrm{PMI}(w_1, w_2) = \log \frac{p(w_1, w_2)}{p(w_1)p(w_2)}$$

where p is the probability of a given occurrence or concurrence and where w_1 and w_2 represent words drawn from the corpus or lexicon.

Semantic Orientation from Pointwise Mutual Information (SO-PMI)

In statistics, Pointwise Mutual Information is a way to measure the association of two outcomes by considering their coincidence (or lack thereof).

$$\mathrm{PMI}(w_1, w_2) = log \frac{\frac{1}{N} hits(w_1 \text{ NEAR } w_2)}{\frac{1}{N} hits(w_1) \frac{1}{N} hits(w_2)}$$

where N is the total number of pages in which the words appear, w_1 and w_2 represent words drawn from the corpus or lexicon, and (w_1 NEAR w_2) means that w_1 is within 10 words of w_2.

Semantic Orientation from Pointwise Mutual Information (SO-PMI)

$$\text{SO-PMI}(w) = \sum_{p \in P} PMI(w, p) - \sum_{n \in N} PMI(w, n)$$

where PMI is as previously defined, p and n are known positive and negative words within the sets P and N, the sets of known positive and negative words.

Semantic Orientation from Pointwise Mutual Information (SO-PMI): Results

Given a large corpus, such as the 100 billion word set of all English-language pages indexed by AltaVista:

- SO-PMI's accuracy vs manual tagging is 87.13% over 100% of the result set.
- SO-PMI's accuracy vs manual tagging is 98.20% over the most confident 25% of its results.

Accuracy is the degree to which the method accurately judges a sentiment to be positive or negative. Confidence is the absolute value, the strength, of the semantic orientation.

Semantic Orientation from Pointwise Mutual Information (SO-PMI): Confounding Factors

Given a smaller corpus, the 10 million word Touchstone Applied Science Associates set:

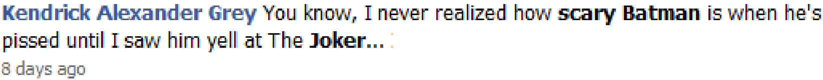
- SO-PMI's accuracy vs manual tagging is 61.26% over 100% of the result set.
- SO-PMI's accuracy vs manual tagging is 47.33% over the most confident 50% of the results.
- SO-PMI's accuracy vs manual tagging is 69.74% over the most confident 25% of its results.

Accuracy and stability suffer when smaller corpora are used.

Semantic Orientation from Pointwise Mutual Information (SO-PMI): Confounding Factors

Why so inaccurate?

- Smaller corpora provide less information.
- Bad movies can have good actors in them.
- Good movies may have fearsome villains or disturbing scenes.



Kendrick Alexander Grey You know, I never realized how **scary Batman** is when he's pissed until I saw him yell at The **Joker**...

8 days ago

Semantic Orientation from Latent Semantic Analysis (SO-LSA)

Latent Semantic Analysis (LSA) is another means of finding the semantic association between a pair of words. LSA uses Term Frequency-Inverse Document Frequency scoring to analyze the statistical relationships between words in a corpus.

$$\text{SO-LSA}(w) = \sum_{p \in P} LSA(w, p) - \sum_{n \in N} LSA(w, n)$$

Semantic Orientation from Latent Semantic Analysis (SO-LSA): Term Frequency

$$tf_{ij} = \frac{n_{ij}}{\Sigma_k n_{ij}}$$

where the numerator, n_{ij}, is the number of times the term (t_i) appears in the document (d_j) and the denominator, $\Sigma_k n_{ij}$, is the total size of the document, the sum of the number of occurrences of every item in the document.

Semantic Orientation from Latent Semantic Analysis (SO-LSA): Inverse Document Frequency

$$idf_i = log \frac{|D|}{|\{d : t_i \in d\}|}$$

where $|D|$ is the number of documents in the corpus and $|\{d : t_i \in d\}|$ represents the number of documents in which $n_{i,j}$ is nonzero. This represents the number of documents in which the word i appears.

Semantic Orientation from Latent Semantic Analysis (SO-LSA): Results

SO-LSA does not yet function on larger corpora, such as the AV-ENG. All results from the use of the TASA set.

- SO-LSA's accuracy vs manual tagging is 65.72% over 100% of the result set.
- SO-LSA's accuracy vs manual tagging is 81.98% over the most confident 25% of its results.

SO-LSA more effectively, stably uses small corpora than SO-PMI. Accuracy is higher and steadily rose with confidence. Accuracy is generally lower compared to SO-PMI over large corpora.

Adjective Conjunction Measurement

The third attempt to increase the reliability of sentiment analysis comes in the form of the analysis of adjective conjunctions. It has been observed that conjunctions imply important information about the orientation of their arguments.

1. "The tax proposal was simple **and** well-received by the public."
2. "The tax proposal was simplistic **but** well-received by the public."

Adjective Conjunction Measurement

The third attempt to increase the reliability of sentiment analysis comes in the form of the analysis of adjective conjunctions. It has been observed that conjunctions imply important information about the orientation of their arguments.

1. “The tax proposal was simple **and** well-received by the public.” - Same Orientation
2. “The tax proposal was simplistic **but** well-received by the public.” - Different Orientation

Adjective Conjunction Measurement

The conjunction-judging system has four main stages.

1. Extracts conjunctions and adjectives from the corpus to make a dictionary of adjective pairs.
2. Combine information from different conjunctions to determine if the adjectives are of similar or different orientation.
3. Assigns each pair of adjectives an “associated dissimilarity value” and attempts to divide the pool of adjectives into groups on that basis.
4. Assigns polarity to the two groups. The group with the highest frequency is positive.

Adjective Conjunction Measurement: Results

The conjunction-judging system has four main stages.

- Average accuracy, when each conjunction is considered independently, is 82%.
- Accuracy can be increased by combining conjunction constraints over multiple pairs of adjectives.
- Accuracy was 92.37% when the average number of links for each adjective was 10.49.
- Accuracy was as low as 78% with lower numbers of links

Natural Language Processing Combined Method

The natural language processing approach augments SO-PMI, is predicated on the idea that the analysis of local statements is more reliable than attempts to judge overall opinion.

1. X admires Y
2. X fails to do Y
3. X provides a good working environment

Natural Language Processing Combined Method

The natural language processing approach augments SO-PMI, is predicated on the idea that the analysis of local statements is more reliable than attempts to judge overall opinion.

1. X admires Y → Y: Positive
2. X fails to do Y → X: Negative
3. X provides a good working environment → X: Positive

Part of speech tagging allows us to understand more of the text.

Natural Language Processing Combined Method: Results

- The Natural Language Processing Combined Method system achieved 94.5% accuracy.
- Accuracy drops to about 75% in some cases involving “well-written texts” such as organizational web pages and news articles.
- “It’s difficult to take a bad picture with this camera.” is seen to be negative: “-1 bad–picture (a bad picture).”

Conclusion

- Natural Language Processing Combined Method was the most effective.
- SO-PMI is the most accurate of the other methods.
- The more a method “understands” the text, the more effective it is.

References

V. Hatzivassiloglou and K. R. McKeown.
Predicting the semantic orientation of adjectives.
In *EACL '97: Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, pages 174–181, Morristown, NJ, USA, 1997. Association for Computational Linguistics.

T. Nasukawa and J. Yi.
Sentiment analysis: capturing favorability using natural language processing.
In *K-CAP '03: Proceedings of the 2nd international conference on Knowledge capture*, pages 70–77, New York, NY, USA, 2003. ACM.

P. D. Turney.
Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews.
In *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 417–424,