AMBASSADOR OF ANALYTICS: Q&A with President-Elect Anne Robinson

informs

ORMS TODAY

December 2012

Volume 39 • Number 6

Sentiment Mining of Social Media

New approach overcomes inherent problems, explores communication channel bias

Also Inside:

- **O.R., ANALYTICS & HEALTHCARE**
  Market-driven reform shifts emphasis from volume to value
- **DATA MINING & FORECASTING**
  Integration, leveraging time-series data offers best model
- **INFORMS' BEST & BRIGHTEST**
  Who won what: Institute presents its achievement awards

Contents

ORMS TODAY

December 2012 • Volume 39 • No.6

Departments

26

Industry experts analyze what's in store for healthcare.

54

Photo essay: Annual meeting draws near-record attendance.

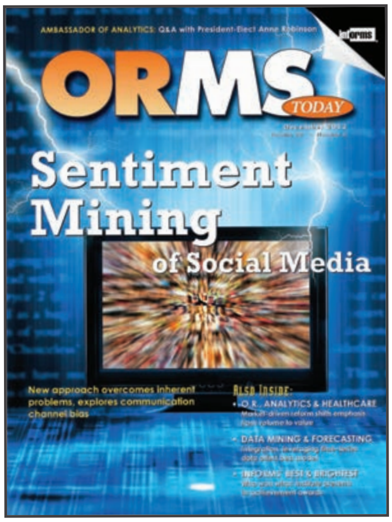
INFORMS News

On the Cover

Mining Media for Meaning

Sentiment analysis explores social media, blogosphere and other communication channels.

Photo © Saniphoto | Dreamstime.com



Features

OR/MS Today (ISSN 1085-1038) is published bimonthly by the Institute for Operations Research and the Management Sciences (INFORMS). Canada Post International Publications Mail (Canadian Distribution) Sales Agreement No. 1220047. Deadlines for contributions: Manuscripts and news items should arrive no later than six weeks prior to the first day of the month of publication. Address correspondence to: Editor, ***OR/MS Today***, 506 Roswell Street, Suite 220, Marietta, GA 30060. The opinions expressed in ***OR/MS Today*** are those of the authors, and do not necessarily reflect the opinions of INFORMS, its officers, Lionheart Publishing Inc. or the editorial staff of ***OR/MS Today***. Membership subscriptions for ***OR/MS Today*** are included in annual dues. INFORMS offers non-member subscriptions to institutions, the rate is $62 USA, $79 Canada & Mexico and $85 all other countries. Single copies can be purchased for $10.50 plus postage. Periodicals postage paid at Hanover, MD, and additional mailing offices. Printed in the United States of America. POSTMASTER: Send address changes to ***OR/MS Today,*** INFORMS-Maryland Office, 7240 Parkway Drive, Suite 300, Hanover, MD 21076. ***OR/MS Today***

© Marco Rullkoetter | Dreamstime.com

Implicit sentiment mining

New approach to sentiment analysis helps overcome inherent problems while exploring media bias and electoral sentiment.

By Maksim Tsvetovat, Jacqueline Kazil and Alex Kouznetsov

Sentiment analysis is a "hot topic" in the analysis of social media, blogosphere and news sources. Marketers, journalists and government officials routinely ask questions along the lines of "What does the Blogosphere think" about a certain subject. In many ways analysis of online sentiment is replacing more labor-intensive methods such as focus groups and polling.

However, Internet sentiment analysis suffers from a number of inherent problems. As a replacement for polling, it has a significant selection bias – results are heavily biased toward large urban centers as the concentration of social media users is significantly higher in urban areas. As a replacement for focus groups, sentiment analysis lacks in-depth inspection of contents of discourse – thus missing the reasons why sentiment tends to take a specific direction.

Sentiment analysis itself is rather inaccurate. Even human coders perform badly in randomized trials, as sentiment is a subjective metric. For example, a phrase "I bought a Honda yesterday" was perceived by 45 percent as sentiment-neutral and 50 percent as positive (i).

Finally, it is rather distasteful that the entire rich, varied, poetic spectrum of human emotions gets reduced to a number between -1 and 1. It is an inelegant approach to a complex problem, throwing out too much good information too early in the process. For example, try this phrase with your favorite sentiment algorithm:

> *"Parting is such sweet sorrow, that I shall say good night till it be morrow."*
>
> ***- William Shakespeare,***
> **"Romeo and Juliet"**

What is Sentiment Analysis?

THE GOAL of sentiment analysis is to determine if a specific passage in text shows positive, negative or neutral sentiment toward the subject. A very large body of literature has been compiled to this extent, including a plethora of papers on NLP [9] and especially analysis of social media and Twitter [4].

In a nutshell, sentiment analysis relies on the fact that words and expressions can be thought of as having emotional meaning and response. Words such as "good," "great" and "amazing" are clearly positive; "bad" and "terrible" are clearly negative; and "cat," "dog," etc. are usually neutral *(ii)*.

These word lists come from a variety of sources, notably WordNet [5] database and the largest electronic thesaurus of English words. On occasion, a corpora of words and emotional responses have been crowdsourced through tools such as Mechanical Turk *(iii)*. While statistical techniques differ, most of this work relies on an a *priori marking* of some words, bigrams or phrases as positive or negative. This type of marking misses a very important point – words that may be construed as neutral or positive in one context can change their emotional connotation if the context changes. For example, in political speech the word "tough" is negative when referring to an economic situation, but becomes positive if the candidate refers to his credentials as a crime-fighter ("tough on crime"). Other words such as "liberal" and "conservative" are considered neutral by most sentiment corpora, but they serve as positive identification markers for the candidates' own constituency and negative slurs for the opponents (e.g., "this candidate is embarrassingly liberal and out of touch").

Sarcasm and double entendre are used daily in online speech – where a single word added to a phrase can completely reverse its meaning (e.g., "Ron Paul wants to legalize marijuana! Oh great!"). Several techniques in the field use a dictionary to mark such words as *inverters or amplifiers (iv)*.

A better, more multi-faceted approach would be to analyze multiple dimensions of sentiment simultaneously. An approach to this idea was proposed by Samsonovic and Ascoli [8]. In their study, emotional content of words in multiple languages has been decomposed into a number of orthogonal dimensions, including:

- ***Valence:*** measures from "good" to "bad" – the traditional sentiment scale.
- ***Nearness or relevance:*** measures from an event in social space ("this is happening to me or my immediate kin" to "somewhere in the world").
- ***Immediacy:*** measures distance from an event in time from "in the past" through "now" to "someday later."
- ***Certainty:*** measures probability of an event occurring, from "definitely" to "most likely not."

Samsonovic and Ascoli identified a number of other dimensions (nine in all), but the principal component analysis shows that most of the variance in human communications that carry sentiment can be described using these top-four dimensions.

Sentiment **analysis** relies on the fact that **words and expressions** can be thought of as having **emotional meaning and response.** Words such as **"good"** and **"great"** are clearly positive; **"bad"** and **"terrible"** are clearly negative; **"cat"** and **"dog"** are usually neutral.

As an example, if your child stumbles and falls, it is an immediate painful event. However, Charlie Chaplin performing a death-defying pratfall is perceived as funny as it is separated from the viewer by a huge social and temporal distance.

The difficulties in this approach are, first and foremost, compiling or crowdsourcing a sizable corpus of words and expressions with their multi-dimensional valence. But also important – and not trivial – is figuring out how these multi-dimensional sentiment vectors can be combined into overall sentiment for larger text passages.

In short, this approach is extremely promising but requires quite a bit more work to be practical.

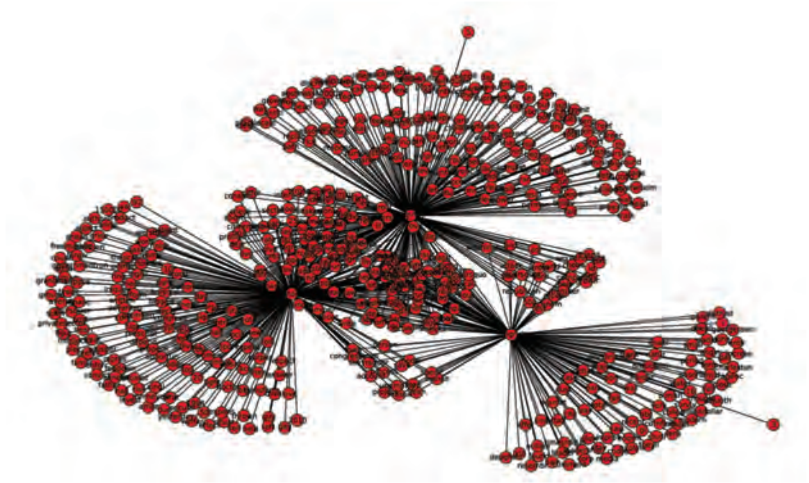


Figure 1: Network of *noun-verb-noun* phrases for three GOP presidential candidates. The phrases in the middle part of the discourse map are shared by all candidates, and the phrases on the periphery are unique speech markers for each of the candidates.

Implicit Sentiment Mining

TO OVERCOME some of the difficulties described above, the co-authors have developed a new approach that sidesteps the issue of sentiment analysis and goes directly to measuring public support for or against public figures.

We rely on a psychological phenomenon called "mirroring" [1, 3]. Mirroring behaviors appear in dialogue between two people when they are interested in each other or identify with each other. Romantic encounters provide perhaps the strongest display of mirroring, as partners unconsciously begin to use same gestures and words and mimic each other's posture and body language. However, similar behaviors appear in weaker social contexts: negotiating parties in business mirror each other, and observers imitate and mirror behavior of pop stars, sports personalities and politicians. Recent studies in social neuroscience show that this effect indeed has biological roots in the parieto-frontal mechanism as a means to allow an individual to understand the action of others "from the inside" and to give the observer a first-person grasp of the motor goals and intentions of other individuals [7].

Similar behavior occurs in speech patterns; strong affinity for a person results in adoption of the person's speech patterns. This happens both offline and, to a slightly lesser but yet noticeable extent, online. The appearance of mirroring in online political speech would thus allow us to understand users' affinities for one or another political candidate, or for and against certain issues. In studying media bias, mirroring will let us determine if certain news sources more readily adopt the language of one or another side of the issue.

Marker Words and Phrases

TO GAUGE mirroring behavior of the public and media, we first need a corpus of text to compare it to:

- Text retrieved from direct speech of Republican primary candidates by parsing their debate performances and speeches as publicized by Associated Press in winter of 2012.
- Text of tweets harvested from across the United States during the same time period.
- Text of 1,500 tweets by the Israeli Defense Forces spokesperson (Twitter: @IDFSpokesperson) and Hamas spokesperson (Twitter: @AlQuassam) during the eight-day armed conflict in November 2012.
- Text of 1,500 tweets each for major news networks, including CNN, FOX News, BBC, Al Jazeera and CNBC.

The text was processed through the following linguistic pipeline:

1. Removal of punctuation, numbers, URLs and Twitter-specific tags (e.g. VIA, RT, MT, etc.).
2. Stop-list application (using MIT English stop-list *(v)*). Any word that is present in the stop-list is dropped from the source text, resulting in a schematic representation of the subjects and objects of the sentence, while removing most of the grammar and sentence structure.
3. Stemming [6] and lemmatization. Stemming and lemmatization bring words to their common forms – for example, the words "stemming," "stemmed," "stemmer" are all turned into the word "stem." In the case of verbs, lemmatization brings verbs into the infinitive form – "was," "were," "will be" are all converted to "be."
4. Parts-of-speech tagging [2] and retrieval of *noun-verb-noun* phrases.
5. Entity tagging (proper name, places, etc.). In this application of the algorithm, we only tag places by using a Yahoo geocoding API. If any of the words in a phrase correspond to a name of a place, they are linked to that place's latitude and longitude – thus making it easier to localize speech. Results of geocoding are cached in a local database, which eventually accumulates many place names, including every possible misspelling and alternative spelling of a place name.

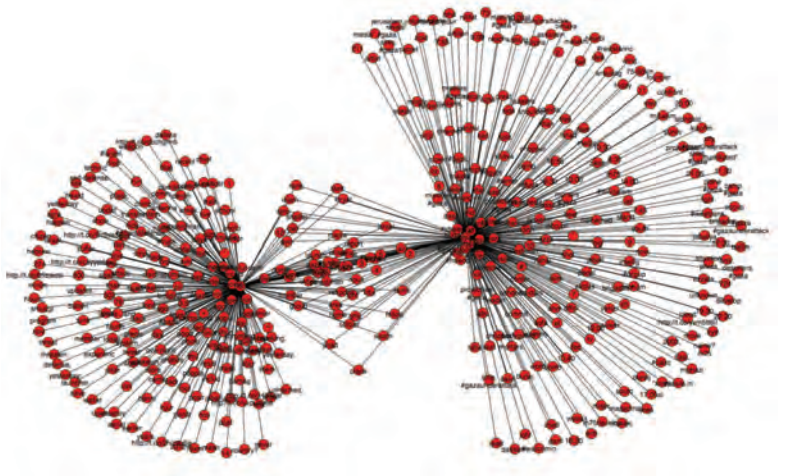


Figure 2: Network of noun-verb-noun phrases for Hamas and IDF Spokesperson. The phrases in the middle part of the discourse map are shared by both parties to the conflict (and mostly include place names where attacks have occurred), and the phrases on the periphery are unique speech markers for each of the parties.

The nouns in *noun-verb-noun* phrases represent nodes in a linguistic network, while the verbs form labels on edges. The resulting network for three GOP presidential contenders (Romney, Gingrigh, Paul) is shown in Figure 1. The center of this network diagram contains terms and entities commonly used by all three candidates – words that are used to legitimize the candidates as (a) American patriots and (b) loyal conservatives.

However, as the primaries are essentially a contest within a single party, the candidates also must distinguish themselves from one another. Words on the fringes of this diagram represent precisely these. We call the words and phrases that are unique for each speaker "marker words."

In the case of Hamas and IDF, the picture is very different. The two spokespeople use different languages (see Figure 2) where the only words in common are names of places that are under attack. The speech of the Hamas spokesperson is rather emotional, using terms such as "the Zionist enemy" and taunting Israeli politicians and media, interspersed with large number of hashtags and URLs. The IDF spokesperson uses drier, more neutral language, usually with no more then one hash-tag and URL per tweet *(vi)*.

The next step in the analysis is to look for these *markers* in the online speech.

Linguistic Mirroring Metrics

WE CREATED A SYSTEM that continuously samples geocoded social media data and processes results through the same linguistic pipeline as outlined above. The only distinction is that in case of social media, we do not attempt to isolate opinions of individual voters but rather gauge aggregate opinion across a geographical area (a county or a state, depending on quality of data). As a result, we generate a two-mode network linking geographic locations with most prominent verb-noun-verb phrases.

The final phase in generating metrics is a comparison of linguistic networks of geographical areas to these of political candidates to find the evidence of mirroring. This is done via a *normalized Hamming distances* between these networks.

Implicit Sentiment in Electoral Data

FOR GOP PRIMARY DATA, the presence of common Republican party markers is used to gauge overall party preference, while presence of individual candidate markers allows us to map individual candidate mirroring behavior and thus propensity to vote for a specific candidate.

Figure 3 shows a map of social media response to Mitt Romney's speech as captured around March 15, 2012. We attempted to predict primary election results for a number of primary contests, including Super Tuesday, where 10 states voted simultaneously. We were able to predict a winner in the densely populated, urban and industrialized states and counties, while mostly rural states received so little data that prediction became nearly impossible. Correcting the rates for state population did not solve the initial problem due to the fact that rural areas have very low social media penetration – it simply highlighted the fact that the urban-rural divide in the United States is real and more pronounced then ever.

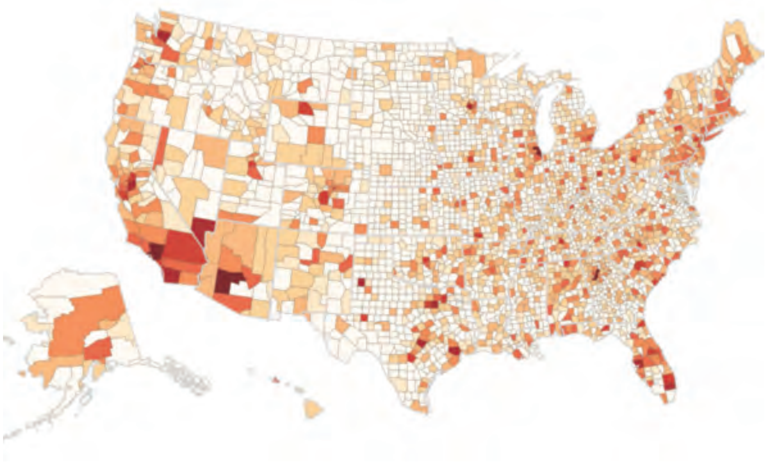


Figure 3: Map of social media response to Mitt Romney in comparison to other GOP candidates, March 2012. Note that data is largely lacking in the center of the country – this is characteristic of all attempts to map online sentiment across the United States and largely correlates to percentage of urban population in each state

Implicit Sentiment and Media Bias

IN CASE OF DATA gathered during the Gaza conflict, the presence of common markers was used to gauge media preferences for using a style of language preferred by one or the other party to the conflict, thus signifying and quantifying media bias.

The tweets originating from major media sources were passed through the same linguistic

One day **traditional** methods, **multi-dimensional** methods and **implicit** methods will be **combined** as an **"emotional processor."**

pipeline as text from the spokespeople. The results of comparison show that a majority of Western media sources generally use language more similar to that of the IDF spokesperson. While it does not indicate direct or indirect support for the cause of Israel vs. the Palestinians, it seems to indicate a level of radicalized speech. In this comparison, the most radicalized of the major networks is FOX News, with CNN not far behind. The least radicalized is CNBC. BBC and Al Jazeera take a more measured approach.

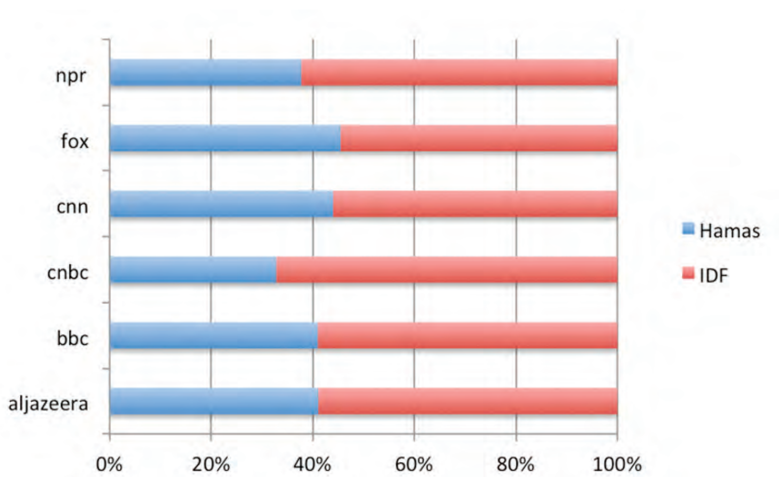


Figure 4: Reflection of IDF and Hamas spokespeople's tweets in major media linguistic signatures.

Conclusions

OUR METHOD IS NOT, by any means, a robust all-around solution to the sentiment-mining problem. Properly solving this would require "hard AI" – a real natural language understanding. However, in this work we are exposing a novel angle to the problem that can be useful in a variety of scenarios, from elections to studying press coverage of major world events. Our hope is that one day traditional methods, multi-dimensional methods and implicit methods will be combined as an "emotional processor" that may not fully understand content of speech, but can understand its emotional content.

Until then ... any dog understands sentiment better then most computers.

Maksim Tsvetovat *(maxt@deepmile.com; Twitter: @maksim2042) is the CTO of DeepMile Networks, a social data analytics consultancy. He also teaches social network analysis at George Mason University. He's the author of "Social Network Analysis for Startups" (O'Reilly), "Hacking Social Complexity" (upcoming, O'Reilly) and more than 50 scientific publications on social network analysis and social data mining.*

Jacqueline Kazil *(jackiekazil@gmail.com; Twitter: @jackiekazil) is a graduate student at George Mason University and a "data lover/pythonista/djangonaut/computational social scientist/founder of PyLadies DC/organizer of @django_district."*

Alex Kouznetsov *(alex@eat-up.org) is an independent open-source software developer working on projects related to mapping and analyzing social data. He is a co-author of the book "Social Network Analysis for Startups," "Hacking Social Complexity" (upcoming) and the primary developer of Agentum, an open-source agent-based modeling toolkit for Python developers.*

Note: Readers interested in replicating the co-authors' results and studying their code can download the Python source code at https://github.com/maksim2042/DC_DataScience_Meetup. The system presented in the open-source repository is a simplified version, with no geocoding and no proprietary code.

REFERENCES

1. H S Baker and M N Baker, 1987, "Heinz Kohut's self psychology: an overview," American Journal of Psychiatry, Vol. 144, No. 1, pp. 1-9, January 1987.
2. Marti A. Hearst, 1992, "Automatic acquisition of hyponyms from large text corpora," in *Proceedings of the 14th International Conference on Computational Linguistics,* pp. 539-545.
3. Steven H Knoblauch, 2009, "From self psychology to selves in relationship: a radical process of micro and macro expansion in conceptual experience," Annals of the New York Academy of Science, Vol. 1159, pp. 262-278, April 2009.
4. Thelwall Mike, Buckley Kevan, Paltoglou Georgios, Cai Di, and Kappas Arvid, 2010, "Sentiment strength detection in short informal text, *Journal of the American Society for Information Science and Technology,* Vol. 61, No. 12, pp. 2,544–2,558.
5. George A. Miller, 1995, "Wordnet: A lexical database for English," *Communications of the ACM,* Vol. 38, No. 11, pp. 39-41.
6. M.F. Porter, 1980, "An algorithm for suffix stripping," *Program,* Vol. 14, No. 3, pp.130-137.
7. Giacomo Rizzolatti and Corrado Sinigaglia, 2010, "The functional role of the parieto-frontal mirror circuit: interpretations and misinterpretations," Nature Reviews Neuroscience, Vol. 11, No. 4, pp. 264-274, April 2010 (print).
8. A. V. Samsonovic and G. A. Ascoli, 2010, "Principal semantic components of language and the measurement of meaning," *PLoS ONE,* Vol. 5:e10921.
9. Peter D. Turney, 2002, "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews," Computing Resource Repository, cs.LG/0212032.

NOTES

(i) http://www.socialmediaexplorer.com/social-media-monitoring/never-trust-sentiment-accuracy-claims/

(ii) In expressions such as "she's a total dog!" or "she's catty" even the innocent pets become emotionally charged. So much for neutrality.

(iii) http://blog.crowdflower.com/2011/11/crowdsourcing-sentiment-analysis-herman-cain/

(iv) http://fjavieralba.com/basic-sentiment-analysis-with-python.html

(v) http://jmlr.csail.mit.edu/papers/volume5/lewis04a/a11-smart-stop-list/english.stop

(vi) These conclusions come from casual observations of following the two accounts for several days, however they seem to be characteristic of the tenor of the discourse.