

Mining the web for the "Voice of the Herd" to Track Stock Market Bubbles

Aaron Gerow¹ & Mark T. Keane²

¹Trinity College Dublin ²University College Dublin

IJCAI-11, Barcelona, Spain

16th – 22nd July 2011



IJCAI

International Joint Conferences on Artificial Intelligence

Can we Predict Financial Crises?

Traditional indicators are often myopic and ineffective.

So...

**Can the language of financial reporting
provide a signal for prediction?**

(cf. Mandlebrot and Shiller, 1963)



The Voice of the Herd

- I. Agreement in financial language is indicative of market conditions**
- II. Temporal changes in verb distributions correlate with market movements**
- III. Affective language shows periods of heightened divergence prior to large shifts**



Method

- Frequency analysis of **lexical** and **syntactic** patterns (verb-phrases)
- Tracking power-law **irregularities**
- Modelling **distributional shifts**
- Tracking distributional shifts in **affective language**



The Corpus

17,716 articles	1 Jan 2006 – 10 Jan 2010
<i>Financial Times</i>	13,286 (ft.com)
<i>New York Times</i>	2,425 (nyt.com)
<i>BBC</i>	2,001 (bbc.co.uk/news)
2006	3,869
2007	4,704
2008	5,044
2009	3,960
2010	136

10.5 M words

300 K sentences

4 M noun-phrases

1.5 M verb-phrases



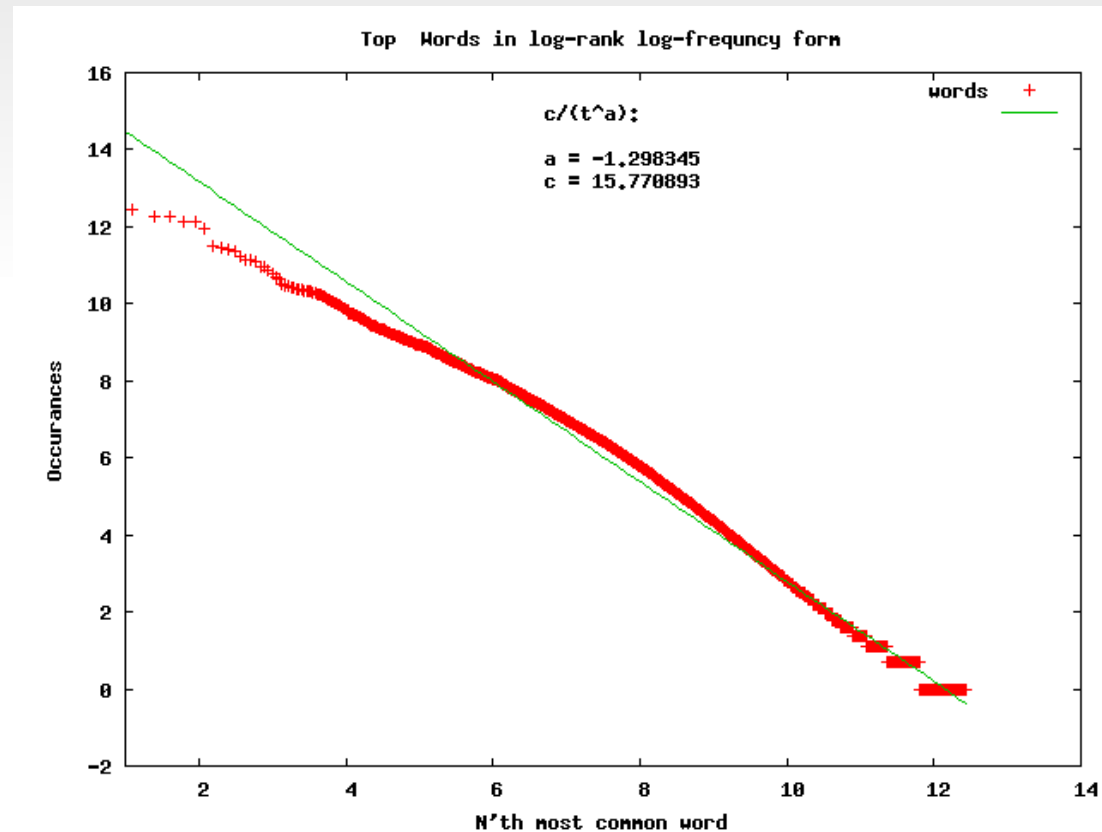
Agreement in Language: Word Distributions

Zipf's law states that word frequencies follow a simple power-law:

$$\forall n \in N, f(n) = \frac{1}{n^a H_N}$$

Zipf's simplified form:

$$f(n) \approx \frac{1}{n^a}$$



[George Zipf]. *Human Behaviour and the Principle of Least Effort*.
Addison-Wesley, Reading, Massachusetts, 1949.



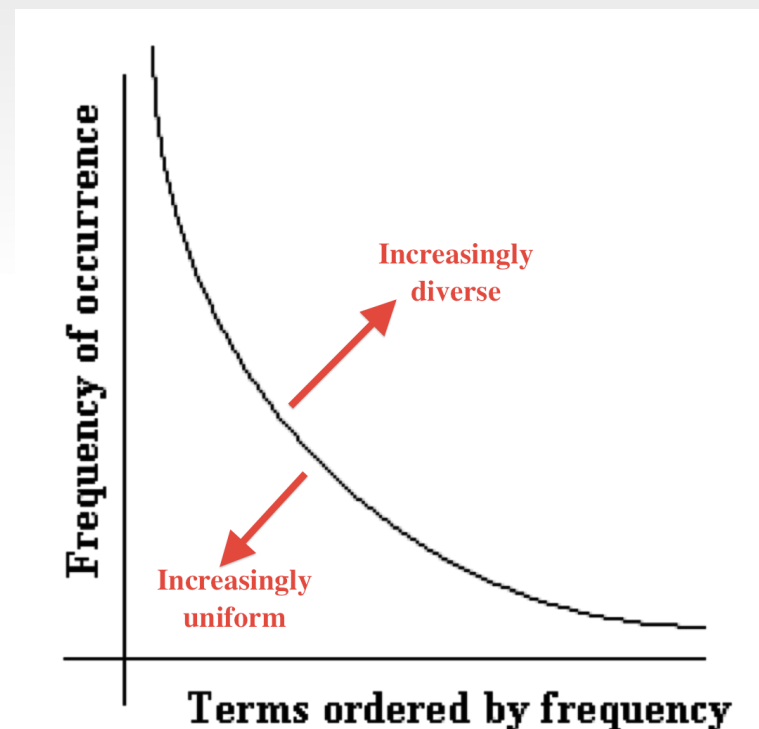
Agreement in Language: Zipf's Law

“Zipf's law states that given some corpus of natural language utterances, the frequency of [linguistic units] is inversely proportional to its rank in the frequency table.”

$$f(n) \approx \frac{1}{n^a}$$

Minor fluctuations, relative to an average, offer a diachronic analysis

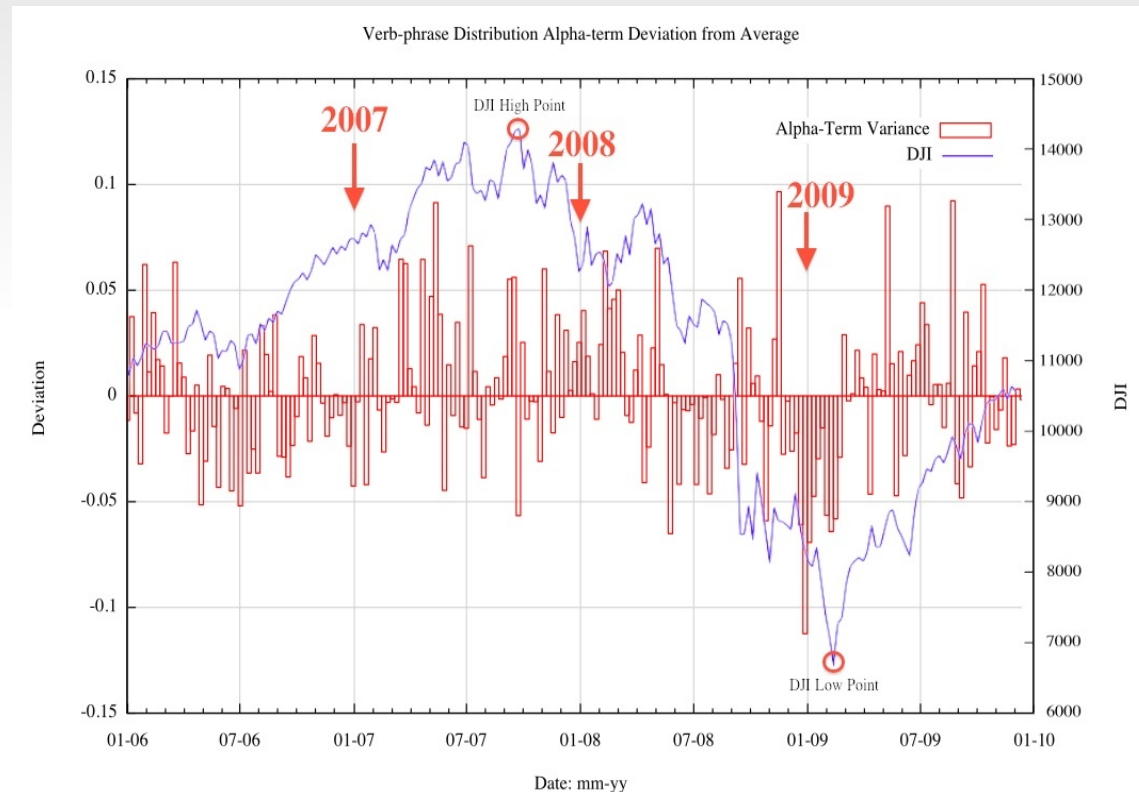
α can parametrise this change



Processing the Corpus

- Articles coded for **source**, **author** and **publication date**.
- Shallow **parsed** (Apple Pie Parser)
- Lemmatised** and tagged (Sketch Engine, TreeTagger)
- Regress for α** in every week's verb-phrase distribution
- Plot weekly **differences** from corpus average

Weekly α , relative to average:



[Satoshi Sekine]. *The Apple Pie Parser*, v5.9. (1997; <http://nlp.cs.nyu.edu/app/>).

[Adam Kilgarriff, Pavel Rychlý, Pavel Smrž, and David Tugwell]. The Sketch Engine. In *Proceedings of EU-RALEX*, pages 105–116, 2004.

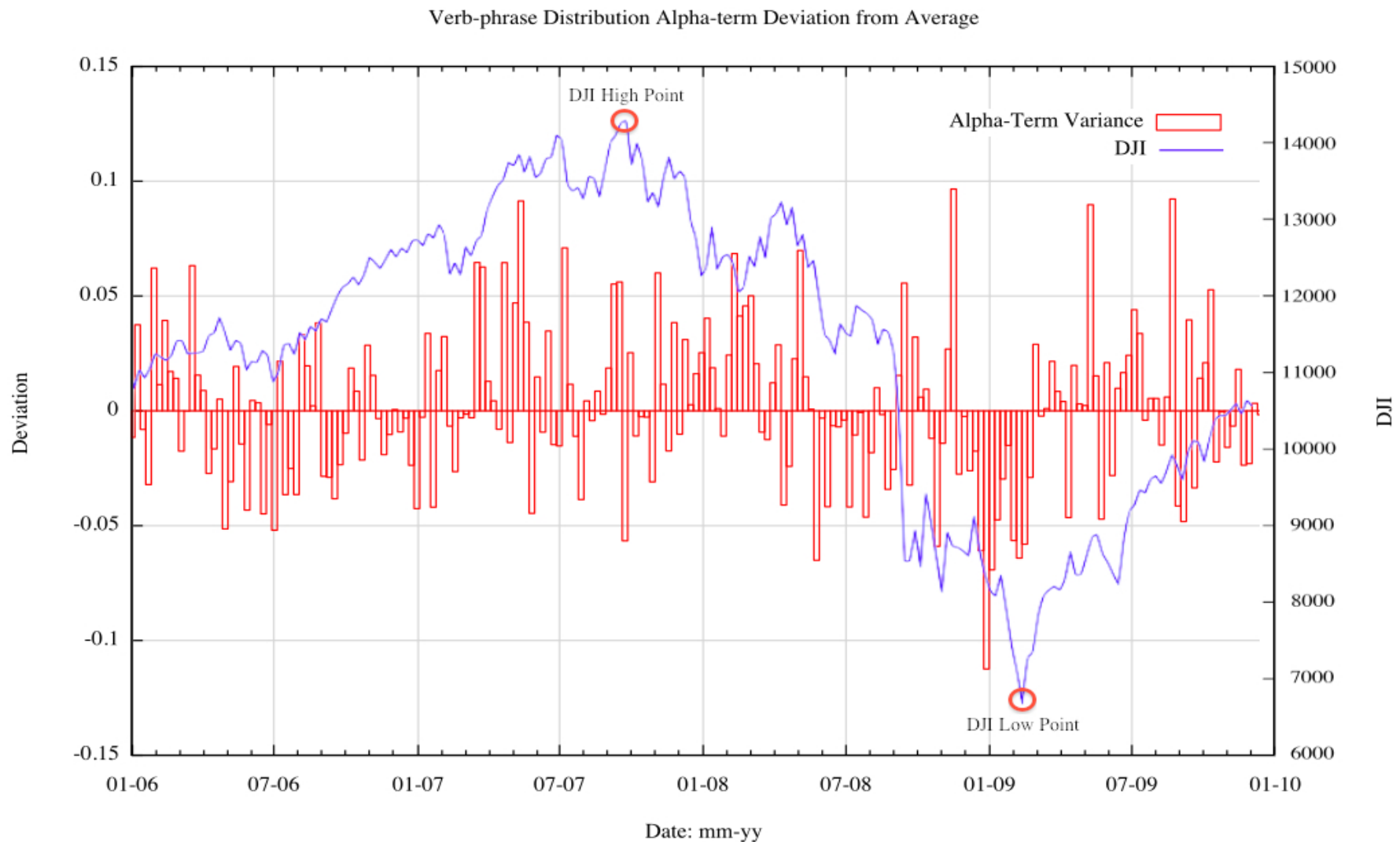


Parts-of-Speech: Why Verbs?

- Nouns encode **subjects** and **objects**
(The tighter the domain, the fewer nouns)
- Adjectives and adverbs encode **qualitative analyses**
- Verbs encode **what is happening**, which is the core of all reporting



Weekly Distributions



The Model

Geometric mean respects non-linearity of residuals:

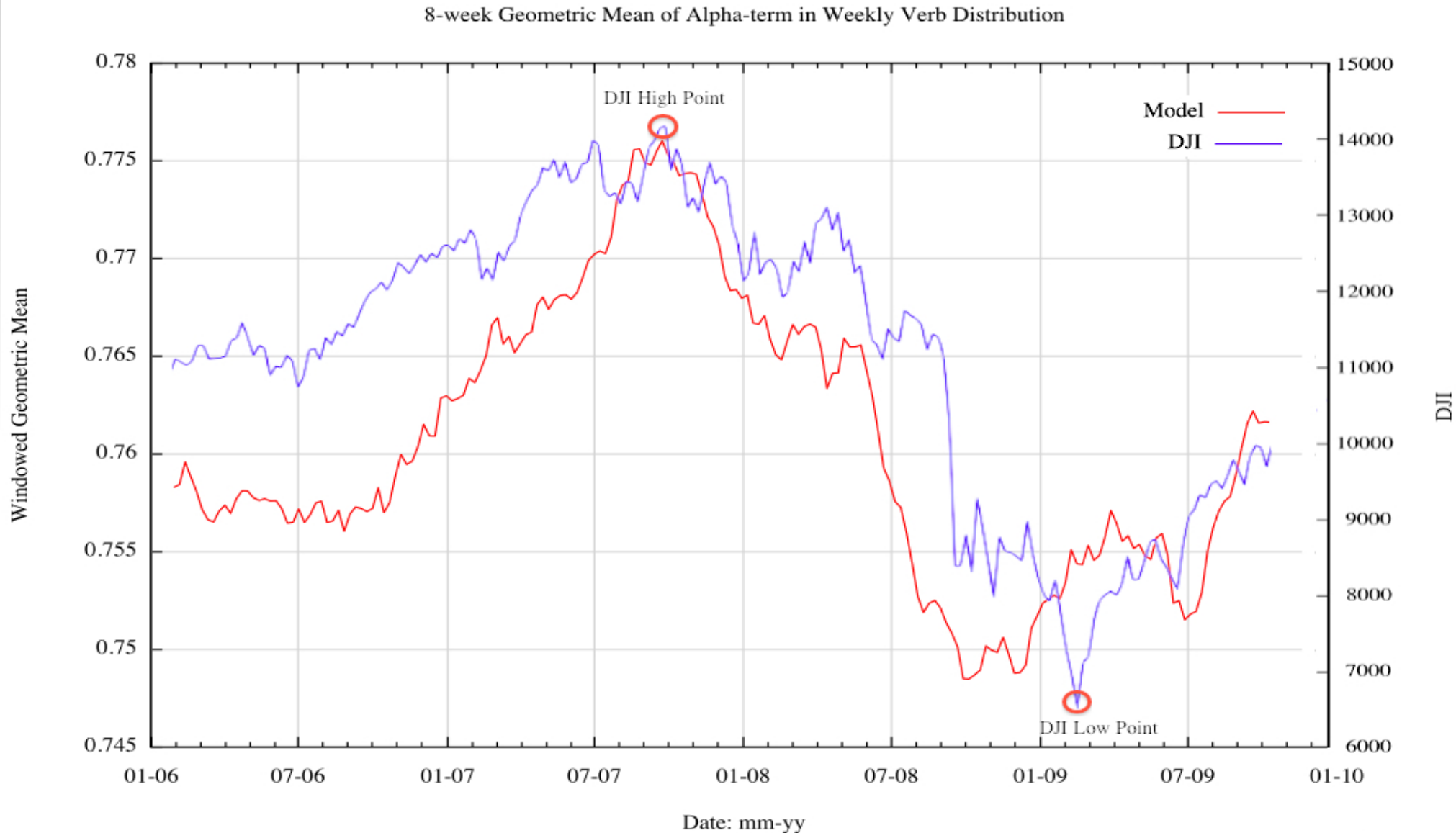
$$\left(\prod_{i=1}^n a_i \right)^{1/n} = \sqrt[n]{a_1 a_2 \cdots a_n}.$$

The model: take a windowed average of α 's

$$week_i = \frac{(\sqrt[8]{a_{i-3} \cdots a_i \cdots a_{i+4}} + \sqrt[8]{a_{i-4} \cdots a_i \cdots a_{i+3}})}{2}$$



The Model



$$r = .79 ; p < .001$$

Conclusions so far

I. Shifts in verb-phrase distribution
characterise the agreement of language

II. These temporal shifts correlate well with
market movements

III. What about another signal:
valency?



Divergence of Affective Language

Verbs tell us what's happening, but can their valence offer a second signal?

Positive

Neutral

Negative



Affective Language Lemma-object Pairs

- Lemmatise the top 200 most common verbs with their most common objects (syntactic).

("stocks rose", "unemployment declined", "Dow Jones fell", etc...)

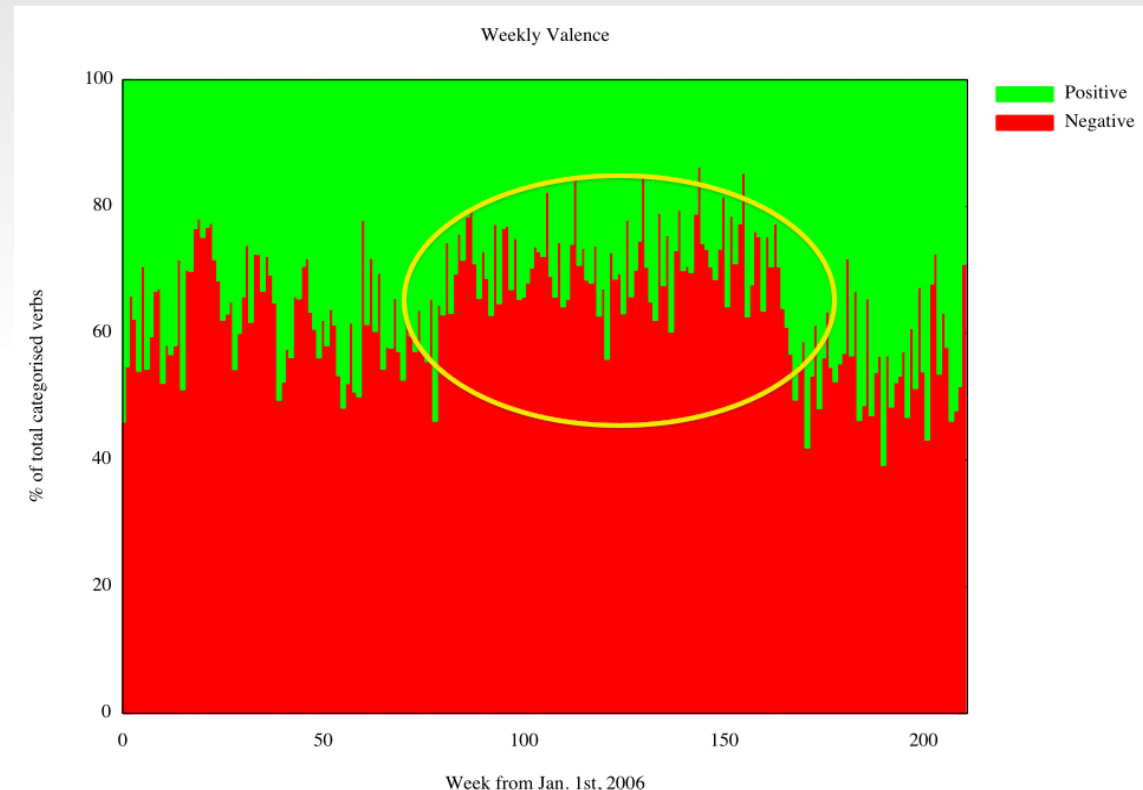
- Hand-classified pairs as positive, negative or neutral / ambiguous:

7,270 positive

14,050 negative

29,590 neutral

Positive / Negative:



Comparing Distributions: Kullback-Leibler Divergence

Traditionally defined as

$$f_{KL}(P||Q) = \sum P_i \log \frac{P_i}{Q_i} \quad \text{or} \quad f_{KL}(P||Q) = \int P_i \log \frac{P_i}{Q_i}$$

Note that

$$f_{KL}(P, Q) \neq f_{KL}(Q, P)$$

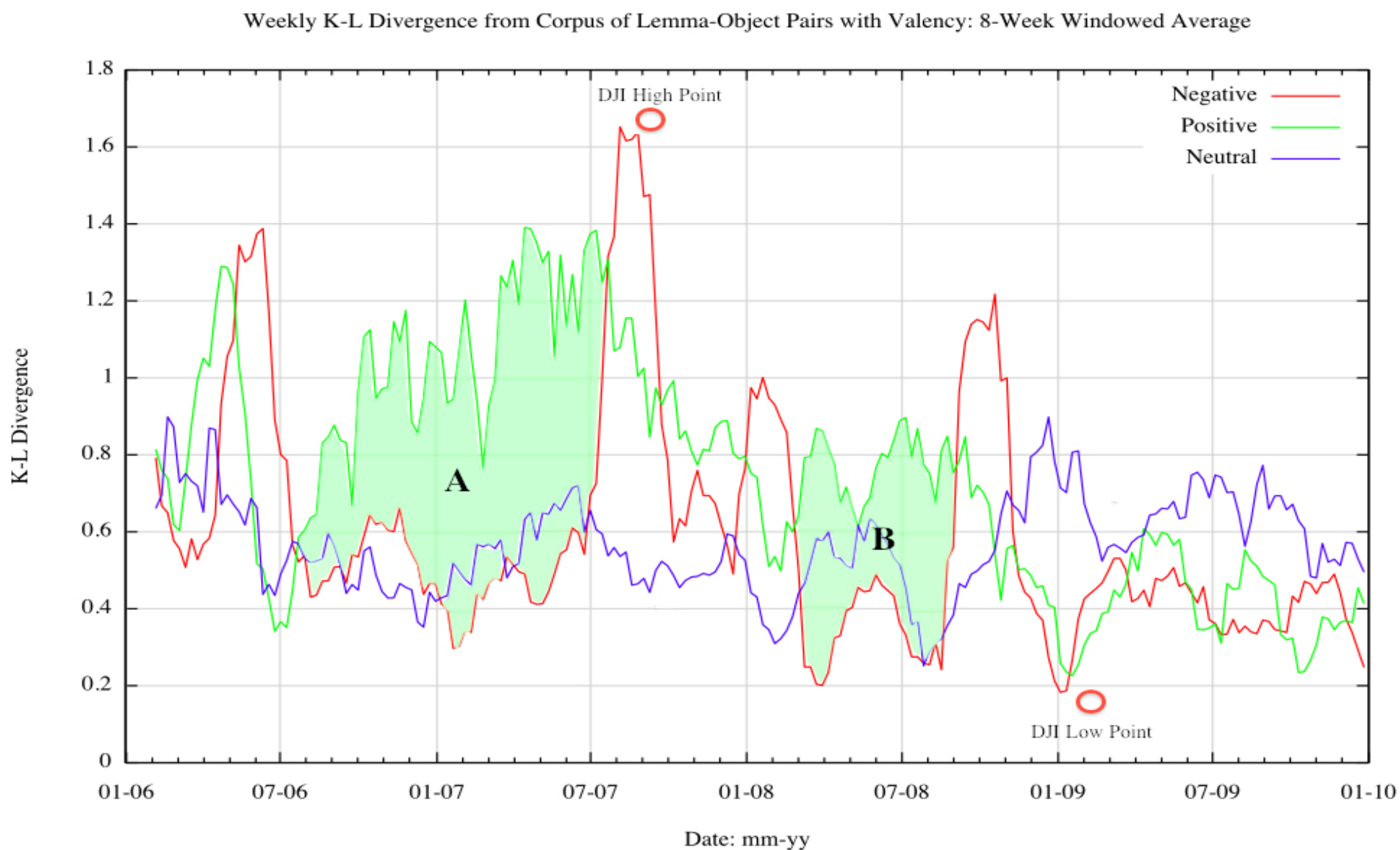
which is easily fixed by defining

$$f_{\text{symetric } KL}(P||Q) = f_{KL}(P, Q) + f_{KL}(Q, P)$$

[Solomon Kullback and Richard A. Leibler]. On information and sufficiency. *Annals Mathematical Statistics* 22: 79–86, 1951.



Divergent Periods



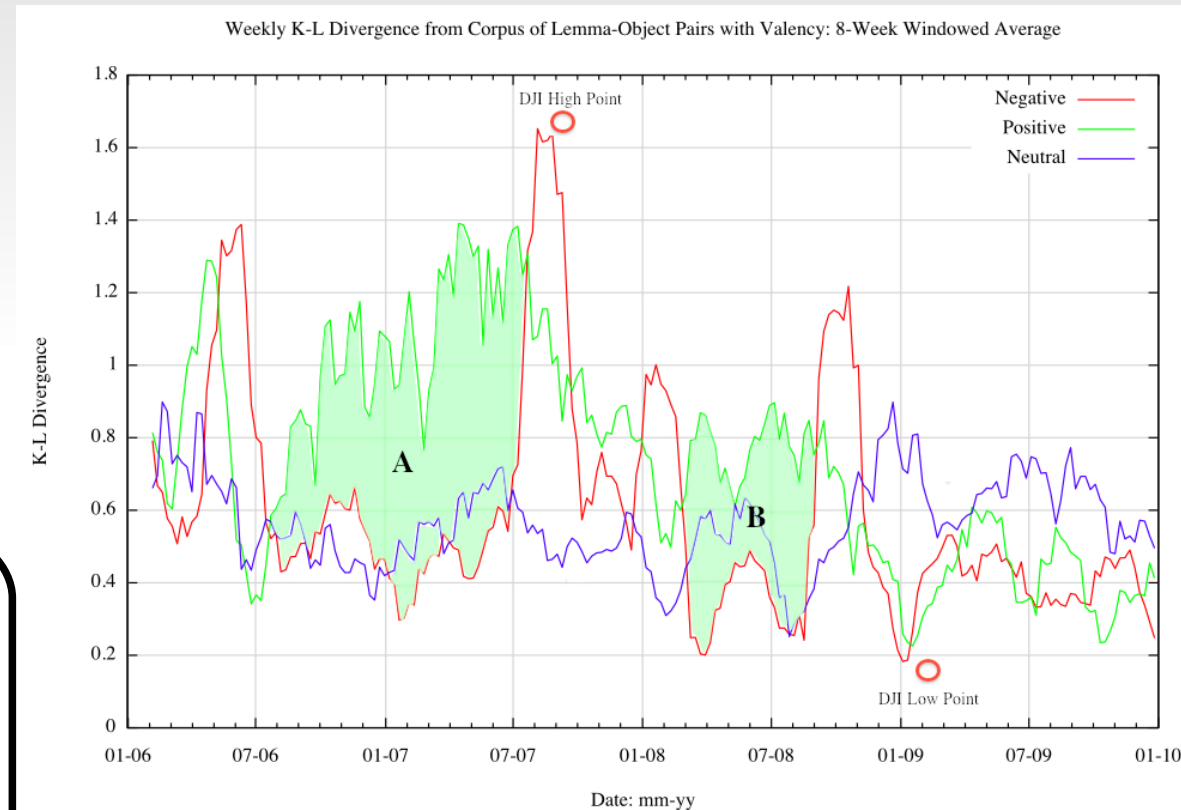
Affective Language: Divergent Periods

Neutral signal is steady

Positive varies in a short
period

Negative signal varies over
larger periods

When positive language
diverges significantly
more than negative,
things are changing



The Voice of the Herd

- I. Agreement in financial language is indicative of market conditions**
- II. Temporal changes in verb distributions correlate with market movements**
- III. Affective language shows periods of heightened divergence prior to large shifts**



Predictors of Financial Crises

Within-system warnings are often explained away by “new valuation models” and “economic paradigms”, which diminish the apprehensions of experts.

Language offers another signal



Thanks to...

- **Toby Walsh, et al. @ IJCAI-11**
- Three anonymous reviewers and

F. Cummins

M. Naughton

N. Stokes

F. Costello

K. Ahmad

K. Hadfield



Mining the web for the "Voice of the Herd" to Track Stock Market Bubbles

Aaron Gerow¹ & Mark T. Keane²

¹Trinity College Dublin ²University College Dublin

IJCAI-11, Barcelona, Spain

16th – 22nd July 2011



IJCAI

International Joint Conferences on Artificial Intelligence

Can we Predict Financial Crises?

Traditional indicators are often myopic and ineffective.

So...

**Can the language of financial reporting
provide a signal for prediction?**

(cf. Mandlebrot and Shiller, 1963)

IJCAI-11, Barcelona - July 2011

Mining the Voice of the Herd



The Voice of the Herd

- I. Agreement in financial language is indicative of market conditions**
- II. Temporal changes in verb distributions correlate with market movements**
- III. Affective language shows periods of heightened divergence prior to large shifts**



Method

- Frequency analysis of **lexical** and **syntactic** patterns (verb-phrases)
- Tracking power-law **irregularities**
- Modelling **distributional shifts**
- Tracking distributional shifts in **affective language**



The Corpus

17,716 articles	1 Jan 2006 – 10 Jan 2010
<i>Financial Times</i>	13,286 (ft.com)
<i>New York Times</i>	2,425 (nyt.com)
<i>BBC</i>	2,001 (bbc.co.uk/news)
2006	3,869
2007	4,704
2008	5,044
2009	3,960
2010	136

10.5 M words
300 K sentences
4 M noun-phrases
1.5 M verb-phrases



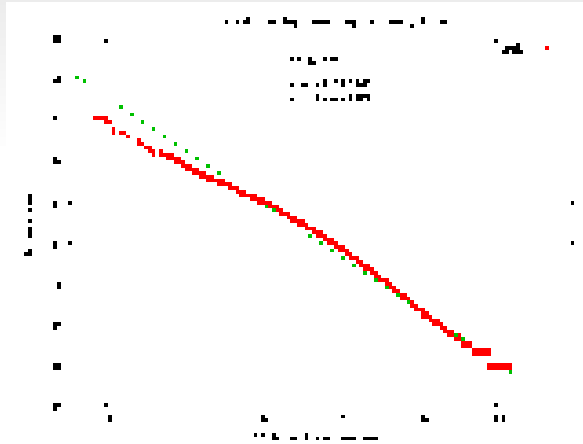
Agreement in Language: Word Distributions

Zipf's law states that word frequencies follow a simple power-law:

$$\forall n \in N, f(n) = \frac{1}{n^a H_N}$$

Zipf's simplified form:

$$f(n) \approx \frac{1}{n^a}$$



[George Zipf]. *Human Behaviour and the Principle of Least Effort*.
Addison-Wesley, Reading, Massachusetts, 1949.

IJCAI-11, Barcelona - July 2011

Mining the Voice of the Herd



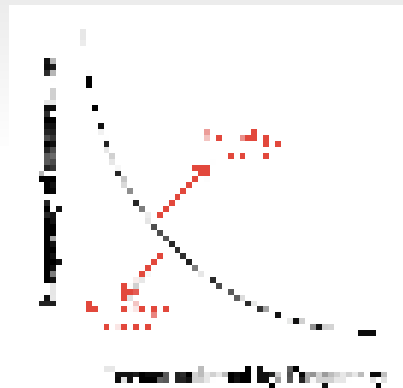
Agreement in Language: Zipf's Law

“Zipf's law states that given some corpus of natural language utterances, the frequency of [linguistic units] is inversely proportional to its rank in the frequency table.”

$$f(n) \approx \frac{1}{n^a}$$

Minor fluctuations, relative to an average, offer a diachronic analysis

α can parametrise this change



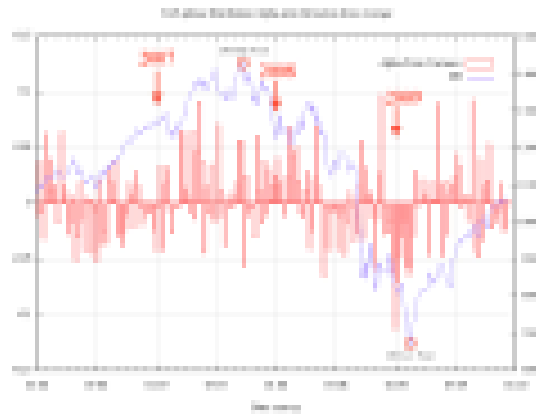
http://en.wikipedia.org/wiki/Zipf's_law



Processing the Corpus

- Articles coded for **source**, **author** and **publication date**.
- Shallow **parsed** (Apple Pie Parser)
- Lemmatised** and tagged (Sketch Engine, TreeTagger)
- Regress for α** in every week's verb-phrase distribution
- Plot weekly **differences** from corpus average

Weekly α , relative to average:



[Satoshi Sekine]. *The Apple Pie Parser*, v5.9. (1997; <http://nlp.cs.nyu.edu/app/>).

[Adam Kilgarriff, Pavel Rychlý, Pavel Smrž, and David Tugwell]. The Sketch Engine. In *Proceedings of EU-RALEX*, pages 105–116, 2004.

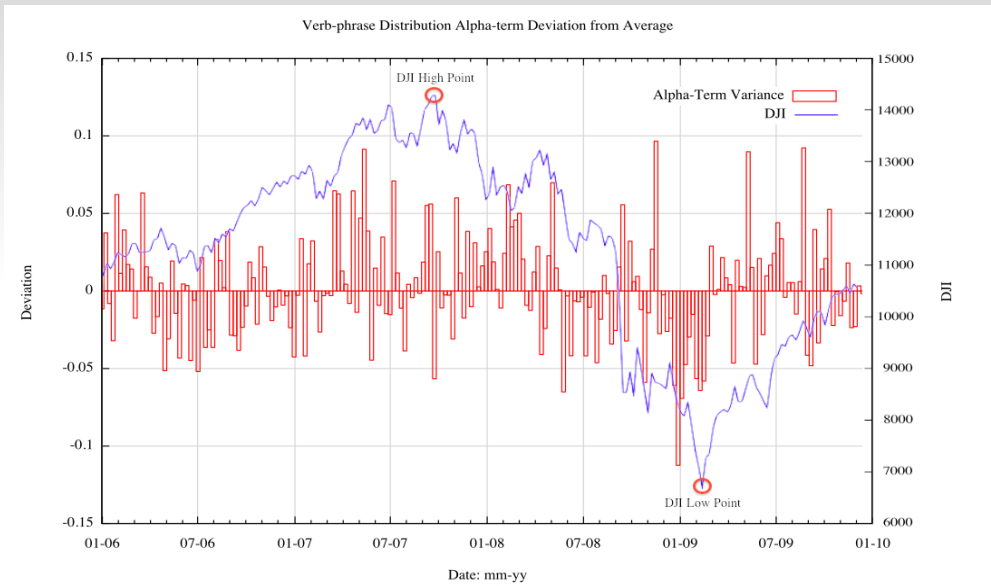


Parts-of-Speech: Why Verbs?

- Nouns encode **subjects** and **objects**
(The tighter the domain, the fewer nouns)
- Adjectives and adverbs encode **qualitative analyses**
- Verbs encode **what is happening**, which is the core of all reporting



Weekly Distributions



The Model

Geometric mean respects non-linearity of residuals:

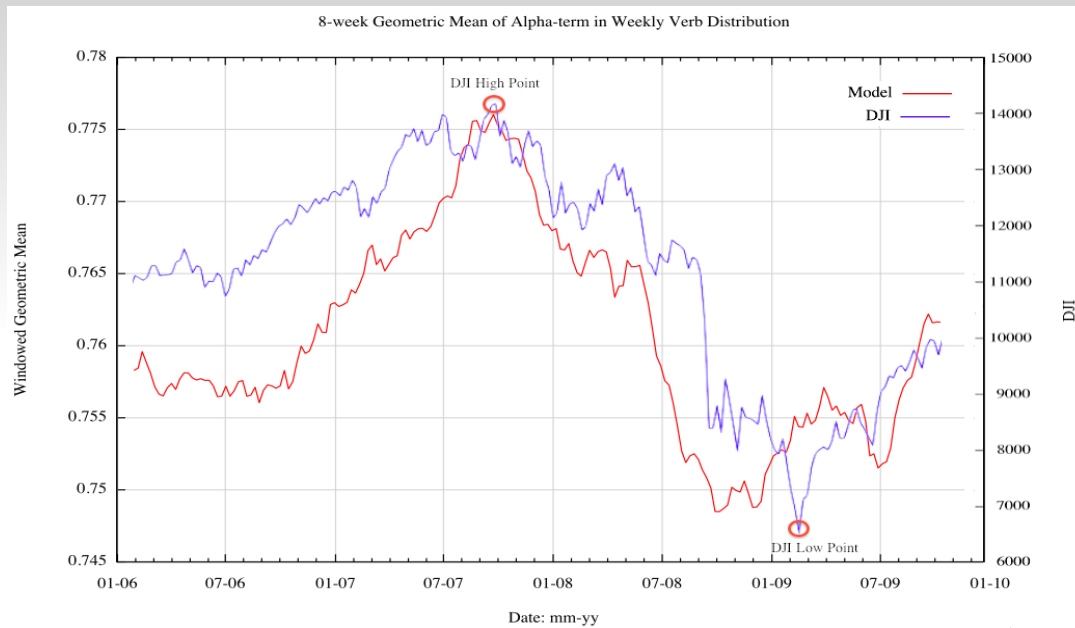
$$\left(\prod_{i=1}^n a_i\right)^{1/n} = \sqrt[n]{a_1 a_2 \cdots a_n}.$$

The model: take a windowed average of α 's

$$week_i = \frac{(\sqrt[8]{a_{i-3} \cdots a_i \cdots a_{i+4}} + \sqrt[8]{a_{i-4} \cdots a_i \cdots a_{i+3}})}{2}$$



The Model



$$r = .79 ; p < .001$$



Conclusions so far

I. Shifts in verb-phrase distribution
characterise the agreement of language

II. These temporal shifts correlate well with
market movements

III. What about another signal:
valency?



Divergence of Affective Language

Verbs tell us what's happening, but can their
valence offer a second signal?

Positive Neutral Negative



Affective Language Lemma-object Pairs

- Lemmatise the top 200 most common verbs with their most common objects (syntactic).

("stocks rose", "unemployment declined", "Dow Jones fell", etc...)

- Hand-classified pairs as positive, negative or neutral / ambiguous:

7,270 positive

14,050 negative

29,590 neutral

Positive / Negative:



42% LOP coverage

This signal, however, concerns frequencies, not distributions like the first analysis. So we'd like a way to compare different distributions to one-another. Cosine similarity, Euclidean distance might spring to mind, but we'd really like something that is easily geared to divergence in a time-series. Kullback and Liebler introduced a "divergence metric" to measure probability the distance between distribution.

Comparing Distributions: Kullback-Leibler Divergence

Traditionally defined as

$$f_{KL}(P||Q) = \sum P_i \log \frac{P_i}{Q_i} \quad \text{or} \quad f_{KL}(P||Q) = \int P_i \log \frac{P_i}{Q_i}$$

Note that

$$f_{KL}(P, Q) \neq f_{KL}(Q, P)$$

which is easily fixed by defining

$$f_{\text{symetric } KL}(P||Q) = f_{KL}(P, Q) + f_{KL}(Q, P)$$

[Solomon Kullback and Richard A. Leibler]. On information and sufficiency. *Annals Mathematical Statistics* 22: 79–86, 1951.

IJCAI-11, Barcelona - July 2011

Mining the Voice of the Herd



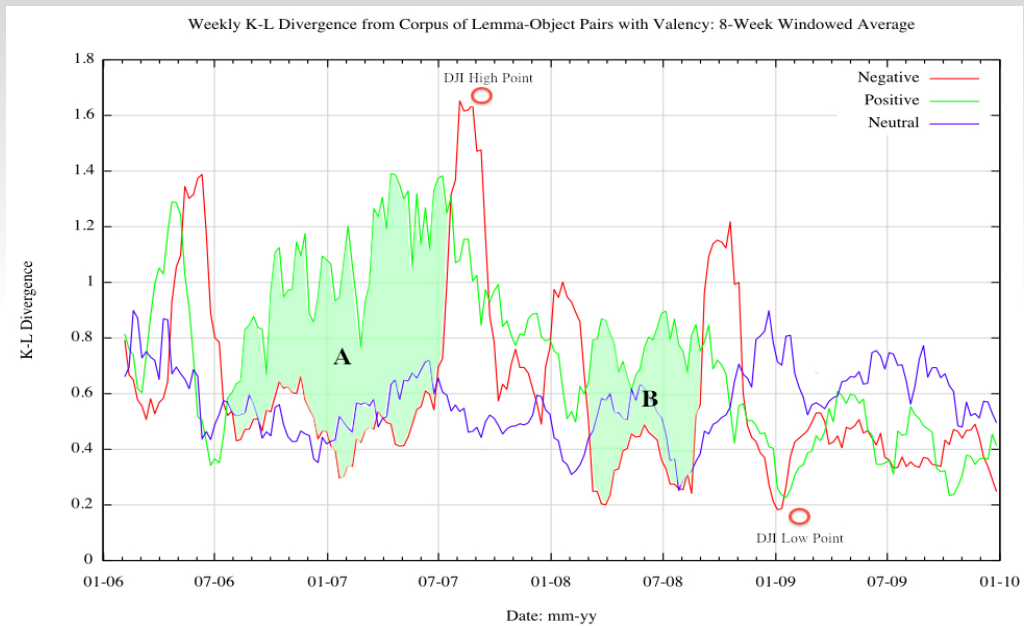
16

Used in NLP to correlate noun-concept shifts with entropy in word distribution to delimit shifts in theme or story.

- Also used in CS to minimise bitwise operations, memory and storage usage, etc...

(Talk about fixing the $Q_i = 0$ case? How /did/ we fix it?)

Divergent Periods



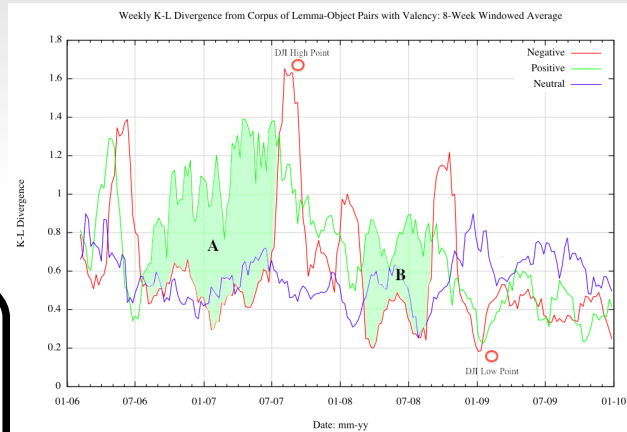
Affective Language: Divergent Periods

Neutral signal is steady

Positive varies in a short
period

Negative signal varies over
larger periods

When positive language
diverges significantly
more than negative,
things are changing



The Voice of the Herd

- I. Agreement in financial language is indicative of market conditions**
- II. Temporal changes in verb distributions correlate with market movements**
- III. Affective language shows periods of heightened divergence prior to large shifts**



Predictors of Financial Crises

Within-system warnings are often explained away by “new valuation models” and “economic paradigms”, which diminish the apprehensions of experts.

Language offers another signal



Paul the Octopus

Thanks to...

- **Toby Walsh**, et al. @ *IJCAI-11*
- Three anonymous reviewers and

F. Cummins
M. Naughton
N. Stokes
F. Costello
K. Ahmad
K. Hadfield

