Seminar 3 — Hedonometer Paper Companion

Reading support for Dodds et al. (2011), *Temporal Patterns of Happiness and Information in a Global Social Network: Hedonometrics and Twitter*

How to use this companion

This document is meant to be read *side-by-side* with the paper. It follows the paper from top to bottom. Whenever you hit a confusing piece of math, a dense paragraph, or a figure that feels hard to "read," come here and look for the matching section, equation, or figure number.

Every key object (a word list, a score, a parameter) is explained without assuming a math background. When the paper uses a formula, we explain what each symbol means and what work the formula is doing.

Throughout, we'll use **"text"** to mean *any pile of words* the authors are analyzing (a day of tweets, an hour of tweets, all tweets containing "BP," etc.).

> Citation for the paper you are reading:
> Peter Sheridan Dodds, Kameron Decker Harris, Isabel M. Kloumann, Catherine A. Bliss, & Christopher M. Danforth (2011). *Temporal Patterns of Happiness and Information in a Global Social Network: Hedonometrics and Twitter*. PLoS ONE 6(12): e26752.

The big idea in one paragraph

The authors build a “hedonometer”: a measuring instrument that assigns a single **average happiness score** to a large body of text (like all tweets posted on a day). They do this by:

1. Building a list of ~10k common words and giving each word a **happiness rating** (from 1 = very unhappy to 9 = very happy) using human judgments on Amazon Mechanical Turk.
2. Counting how often each word appears in a target text.
3. Taking a **frequency-weighted average** of word happiness scores.

They then use this instrument to study *patterns over time* (daily cycles, weekly cycles, special events) and compare the happiness signal to an **information / diversity** signal (how varied the vocabulary is).

Minimal glossary of symbols and terms

Words, counts, probabilities

- w_i: the *i-th word type* in the authors’ word list (e.g., “love,” “war,” “pizza”).
- f_i: the **raw count** of how many times w_i appears in a text T.
- p_i: the **normalized frequency** (a probability-like number):

$$p_i = \frac{f_i}{\sum_j f_j}$$

  Interpretation: “What fraction of all counted words are of type w_i?”

  Key property: $\sum_i p_i = 1$.

Happiness

- $h_{avg}(w_i)$: the **average happiness rating** of word w_i from Mechanical Turk (a number between 1 and 9).
- $h_{avg}(T)$: the **average happiness score of a text** T (computed from word frequencies and word ratings).
- Δh_{avg}: the size of the *neutral zone* around 5 that the authors remove to reduce noise.
  - If $\Delta h_{avg} = 1$, then words with happiness between 4 and 6 are treated as “neutral” and excluded from the hedonometer calculation.

Important: these are **not** the usual grammatical stopwords (“the,” “and,” “of”). Here, “stop words” means *emotionally neutral words*.

Word shift graphs (how they explain changes)

When comparing two texts:

- T_{ref} (“reference”): the baseline text you compare against (e.g., the 14-day window around an event).
- T_{comp} (“comparison”): the special text you care about (e.g., the day of the event).

The shift graph breaks the happiness difference into per-word contributions.

Word diversity / information

- S: **Simpson’s concentration**, defined as

  $$S = \sum_{i} p_{i}^{2}$$

  Interpretation: probability that *two randomly chosen word tokens* from the text are the **same**.
- N_S: **Simpson lexical size**, defined as

  $$N_S = \frac{1}{S}$$

  Interpretation: the “effective number of equally common words.” If the text behaved as if it had N_S equally frequent words, it would have the same concentration.

Abstract and opening framing

1. **Claim: tweets as "in-the-moment" language.** The authors claim Twitter is tuned to *experiential* (moment-to-moment) happiness rather than reflective life evaluation.
2. **Claim: scale fixes noisiness.** Their method is too crude for a sentence, but becomes meaningful when you average across millions of words.
3. **Claim: interpretability.** Because they can say which words pushed the score up/down (word shift graphs), they argue their measure is not a black box.

If you're reading as a humanist: notice the *rhetorical work* of the "measuring instrument" metaphor. The authors want you to see this as comparable to a thermometer: imperfect, but calibrated and useful.

Section 1 — Description of the data set

What is their data, in concrete terms?

They collected tweets over about three years (Sept 2008–Sept 2011), totaling **billions of tweets** and **tens of billions of words**.

- For an early portion of the period, the authors estimate they collected on the order of **~8%** of all tweets posted up to that point; later platform changes made this harder to estimate cleanly, but they argue they still captured **>5%** overall.
- A key metadata change happens on **May 21, 2009**: Twitter began reporting *local time* (instead of GMT) in a way the authors could use for hour-by-hour "daily cycle" analysis.

Important nuance: their feed was not constant. The sampling rate changes over time because Twitter changed its systems and because of outages. The authors emphasize this because it affects how you interpret time series (especially long trends).

What counts as a "word" for them?

A "word" is basically: "a contiguous set of characters separated by whitespace or some punctuation." They do **case-insensitive** matching and they do **not** stem words (so "love" and "loved" are different).

This is a deliberate choice: it makes the method transparent and fast, but it also means:

- misspellings count as their own "words"
- URLs and usernames can count as "words"
- language mixing is unavoidable

Figure 1 (word usage through the day): how to read it

Figure 1 shows that the authors can detect *very ordinary daily rhythms* in the language:

- Words like "breakfast," "lunch," "dinner" peak at plausible hours.

- Other food words have their own rhythms.

Why include this figure so early? It's doing two things:

1. It reassures you that the dataset has a meaningful time signal (not random noise).
2. It sets up the claim: if daily life rhythms are visible in word frequencies, *maybe emotional rhythms are too.*

Section 2 — A robust method for measuring emotional content

This is the mathematical heart of the paper. The good news is: it's essentially a **weighted average**.

2.1 Algorithm for the Hedonometer

Equation (1): the happiness of a text

The key formula is:

$$h_{avg}(T) = \frac{\sum_{i=1}^{N} h_{avg}\left(w_i\right) f_i}{\sum_{i=1}^{N} f_i} = \sum_{i=1}^{N} h_{avg}\left(w_i\right) p_i$$

Translated it into English:

- Look at each word w_i in your word list.
- Take its happiness score $h_{avg}\left(w_i\right)$.
- Multiply by how common it is in the text (p_i or f_i).
- Add these up across words.

The second version (using p_i) is the cleaner idea: it's a *weighted average* where the weights are "how much this word shows up."

A tiny toy example

Suppose your text is:

> "love love hate"

Assume:

- $h_{avg}(love) = 8$
- $h_{avg}(hate) = 2$

Counts:

- $f_{love} = 2$
- $f_{hate} = 1$

Probabilities:

- $p_{love} = 2/3$
- $p_{hate} = 1/3$

Now compute:

$$h_{avg}(T) = 8 \cdot (2/3) + 2 \cdot (1/3) = 16/3 + 2/3 = 18/3 = 6.$$

So the text is "somewhat happy" because it contains *more* "love" than "hate."

What this formula assumes (and what it ignores)

This formula treats text as a **bag of words**:

- Word order does not matter.
- Negation is not handled ("not happy" still contains "happy").
- Sarcasm is not handled.
- Multi-word meaning ("sick" meaning "good") is not handled.

The authors accept these limitations because they are analyzing **very large collections** of tweets, not single sentences.

> Connection to Lecture 3: This is the same move as turning a document into a **vector of word counts** and then taking a weighted average (like a dot product).

2.2 Word evaluations using Mechanical Turk

Where do the word happiness scores come from?

They build a word list (labMT 1.0) and ask many people to rate each word on a 1–9 happiness scale. The result is:

- **average happiness** for each word
- **standard deviation** (how much raters disagree)

Each word is rated by **50** independent respondents.

Where does the word list come from?

The list is not "all English words." It is built from **frequency**:

- Take the 5,000 most frequent words from each of four corpora:
  - Twitter
  - Google Books

    - The New York Times
    - Music lyrics
- Take the union (remove duplicates) → about 10,222 unique word types.

Interpretation: labMT is trying to cover "common words you actually see" rather than rare dictionary entries.

Data Set S1: what is in it?

The supplementary file has 8 columns:

1. word
2. rank (in the labMT list)
3. average happiness (mean of 50 ratings)
4. standard deviation of happiness
5. Twitter rank (frequency rank in top 5000)
6. Google Books rank (frequency rank in top 5000)
7. New York Times rank (frequency rank in top 5000)
8. Music lyrics rank (frequency rank in top 5000)

A double dash "–" means the word is not in the top 5000 for that corpus.

What a humanist should notice here

This dataset is already a sequence of interpretive choices:

- which corpora count as "language"
- which words are "worth rating" (only high-frequency ones)
- whose ratings count (Mechanical Turk workers, in 2011)
- the meaning of "happiness" as a 1–9 number

The dataset looks "objective," but it is built from many social decisions.

2.3 Robustness and Refinement (why they drop "neutral" words)

The central problem they are trying to solve

Many very frequent words are emotionally neutral (think "the," "and," but also words like "okay," "been," "home," etc.). If you include lots of neutral words, they can "drown out" the emotional signal and reduce sensitivity.

So the authors remove words near neutral happiness = 5.

The parameter Δh_{avg}

They define a neutral band around 5:

- Keep word w only if $h_{avg}(w) \leq 5 - \Delta h_{avg}$ or $h_{avg}(w) \geq 5 + \Delta h_{avg}$.

So:

- $\Delta h_{avg} = 0$ keeps everything.
- $\Delta h_{avg} = 1$ removes words between 4 and 6.
- Larger Δh_{avg} removes more words (only very emotional words remain).

They choose $\Delta h_{avg} = 1$ as a compromise. In the paper's Twitter analysis, that leaves **3,686** words and captures about **22.7%** of the total word tokens.

Figure 2: how to read the multi-panel argument

Figure 2 is the authors' "calibration report." It is showing you that the hedonometer is both:

- **robust** (doesn't collapse if you change details)
- **tunable** (you can adjust it like a lens)

Panel-by-panel:

- **2A (time series under different Δh_{avg})**

  As you increase Δh_{avg} (remove more neutral words), the curve wiggles more.

  Interpretation: with fewer words, each remaining word has more influence → higher sensitivity, but also more volatility.

- **2B (distribution of word happiness ratings)**
  The word ratings are not centered at 5; they skew positive.
  Interpretation: English (as measured here) shows a "positivity bias": more common words are rated slightly positive.

- **2C (correlation matrix across Δh_{avg} choices)**

  Even when you change Δh_{avg}, the time series are still strongly correlated.

  Interpretation: the *overall pattern* is stable even though the amplitude changes.

- **2D (number of words remaining as a function of Δh_{avg})**

  As the neutral band widens, the word list shrinks fast.

- **2E (coverage of Twitter tokens vs Δh_{avg})**

  Removing neutral words reduces how much of the Twitter text you are measuring (because you are ignoring many tokens).

- **2F (coverage vs word rank)**
  This shows how much coverage you get from more frequent words.
  Interpretation: because word frequencies are very unequal, the first few hundred words carry a lot of "mass."

The story Figure 2 tells: **you can trade coverage for emotional "contrast."** The authors decide $\Delta h_{avg} = 1$ is a reasonable compromise.

2.4 Limitations

1. **Small text is unreliable.** One tweet is too short; they rely on aggregation.
2. **They measure "exhibited tone," not inner feeling.** This is about public expression, not private emotional states.
3. **They deliberately avoid "deep NLP."** Their method is transparent but crude: it ignores grammar and context.
4. **Gaming and platform incentives exist.** People (or bots) can manipulate expression systems.

Section 3 — Measuring word diversity (information content)

The paper is doing two parallel measurements:

- Happiness (h_{avg})
- Word diversity (N_S)

The diversity measure matters because it can reveal *language mixture* (English vs Spanish) and topical variety.

The basic building block: generalized entropy

They define a generalized entropy-like quantity:

$$J_q = \sum_i p_i^q$$

- p_i is the normalized frequency of word i.
- q controls sensitivity:
  - big q pays more attention to common words
  - small q pays more attention to rare words

They then convert these into an "equivalent lexical size": the number of equally common words that would give the same measure.

Why they choose Simpson lexical size (N_S)

They focus on $q = 2$, which yields **Simpson's concentration**:

$$S = \sum_i p_i^2$$

and then define:

$$N_S = \frac{1}{S}.$$

Interpretations:

- S is the probability that two randomly chosen words are the same.
- N_S is the effective vocabulary size (“how many unique words would we need to reproduce this level of lexical diversity”).

Methods note: why only 50,000 words?

Because Twitter has an enormous number of unique strings (misspellings, URLs, etc.), they restrict diversity calculations to the **50,000 most frequent words** and show this is sufficient for accurate estimates for $q \geq 1.5$ (including $q = 2$).

The word shift graphs (central interpretive tool)

You will see shift graphs repeatedly (Figures 4 and 8 especially). This section explains how to “read” them.

Equation (2): difference between two texts as a sum over words

They start by writing:

$$h_{avg}^{(comp)} - h_{avg}^{(ref)} = \sum_{i=1}^{N} h_{avg}(w_i)\left[p_i^{(comp)} - p_i^{(ref)}\right]$$

Then they rewrite it to make interpretation easier:

$$h_{avg}^{(comp)} - h_{avg}^{(ref)} = \sum_{i=1}^{N} \left[h_{avg}(w_i) - h_{avg}^{(ref)}\right]\left[p_i^{(comp)} - p_i^{(ref)}\right].$$

Read the two brackets as two questions:

1. **Is this word happier or sadder than the reference average?** $\left[h_{avg}(w_i) - h_{avg}^{(ref)}\right]$
2. **Did this word become more or less common in the comparison text?** $\left[p_i^{(comp)} - p_i^{(ref)}\right]$

Multiply those two answers and you get the word’s contribution.

Equation (3): converting contributions to percentages

To make plots readable, they normalize contributions so they sum to 100%:

$$\delta h_{avg,i} = 100 \times \frac{\left[h_{avg}(w_i) - h_{avg}^{(ref)}\right]\left[p_i^{(comp)} - p_i^{(ref)}\right]}{\left|h_{avg}^{(comp)} - h_{avg}^{(ref)}\right|}.$$

This is just saying: “How much of the total change is due to word i?”

How to actually read a word shift figure

A practical method:

1. Look at the title: what are T_{ref} and T_{comp}?
2. Note whether the day is overall happier or sadder than the reference.
3. Read the top ~10 words and their directions:
   - arrows indicate whether frequency went up or down
   - The +/- indicates whether the word is positive or negative relative to the reference average
4. Use the bottom-right “four circles” summary to see the *type* of change:
   - Is it mostly “more negative words”?
   - Or mostly “less positive words”?
   - Or a mix?

Results and Discussion — what each section is doing

Section 4 — Overall time dynamics of happiness and information

Figure 3: what you are seeing

Figure 3 has three aligned time series:

- **3A**: daily average happiness of all tweets
- **3B**: daily Simpson lexical size (N_S)
- **3C**: the number of word tokens that were usable for happiness (i.e., found in labMT with $\Delta h_{avg} = 1$)

Important reading point: **the happiness values move in a narrow band** (roughly 5.9–6.2). The authors treat small shifts as meaningful only because the dataset is huge.

Section 4.1 (Outlier dates): what counts as an “event”

They mark dates where the happiness score sharply deviates from nearby days. They argue:

- Positive outliers are mostly predictable cultural holidays (Christmas, New Year’s, etc.)
- Negative outliers often follow unexpected collective trauma (disasters, deaths, etc.)

As a reader, notice that this is a *storytelling move*: they connect spikes to recognizable events to validate the instrument.

Figure 4: three worked examples of word shifts

Figure 4 is showing *why* the time series spikes happen, using word shifts for:

- U.S. financial bailout (negative)
- Royal Wedding (positive)
- Death of Osama bin Laden (mixed / complex)

What to focus on:

- The *event-specific words* rise ("bailout," "wedding," "dead," etc.)
- The shift is often dominated by surprisingly few words: the first ~1000 words explain almost all of the change.

Section 4.3 (Information content): why N_S rises over time

The authors observe a strong increase in N_S (effective vocabulary size) starting mid-2009. They explain this as:

- An increase in non-English tweeting (especially Spanish common words like "que," "la," "y," etc.)

This is a key interpretive lesson: *word diversity measures can detect language mixture and platform shifts.*

Section 5 — Weekly cycle

Figure 5: the "average of averages" trick

They want a clean weekly pattern, but tweet volume changes over time. So they compute:

- happiness for each each day separately
- then average those daily score scores

This gives each day of the week equal weight (rather than letting high-volume Mondays dominate).

Figure 6: "Monday" is literally rated as less happy

They show Mechanical Turk ratings for the words "Monday," "Tuesday," etc. Saturday and Sunday are rated much happier than Monday.

Interpretation: part of the weekly cycle is baked into the emotional associations of day-names themselves.

Figure 7: stability over time

They repeat the weekly pattern across four time windows and show it is stable. The point is: this is not a one-time artifact.

Figure 8: Saturday vs Tuesday word shift

Figure 8 explains “happier Saturdays” in terms of word use:

- more leisure / social words (positive)
- fewer work / stress words (negative)
- but also some increases in negative weekend words (“hangover,” “drunk,” “fight”)

This is a good example of the method producing a nuanced picture: weekends are happier *and* messier.

Figure 9: weekly pattern in word diversity is different

The Simpson lexical size peaks on Friday and dips on Sunday. The authors interpret this as shifts in conversational focus and predictability.

Section 6 — Daily cycle

Figure 10: average happiness by hour (local time)

They find:

- lowest happiness around 10–11 pm
- highest happiness around 5–6 am (“biological midnight” idea)

The variation across a day is stronger than across a week. The authors compare this to other mood studies (blogs, surveys) and note differences.

Figure 11: profanities track the opposite pattern

They show common profanity usage peaking around 1 am and being lowest when happiness is highest. This supports (but does not prove) the mood-cycle interpretation.

Figure 12: morning vs night word shift

This is a key “how to interpret” figure:

- morning is happier partly because negative words are less frequent
- and partly because certain positive words are more frequent

The authors emphasize that the shift is dominated by a relatively small set of words.

Figure 13: daily cycle in word diversity

They find N_S peaks around 5–6 am and bottoms around 10–11 pm (similar shape to happiness). They interpret parts of this via changes in common function words and interaction patterns (pronouns, retweets, etc.).

The diversity-shift equations (4) and (5)

For lexical size, they can also decompose differences into word contributions.

They start from $N_S = 1/S$ and derive:

$$N_S^{(comp)} - N_S^{(ref)} = \frac{1}{S^{(comp)}S^{(ref)}} \sum_{i=1}^{N} \left[\left(p_i^{(ref)}\right)^2 - \left(p_i^{(comp)}\right)^2 \right]$$

and define per-word percent contributions:

$$\delta N_{S,i} = 100 \times \frac{\left(p_i^{(ref)}\right)^2 - \left(p_i^{(comp)}\right)^2}{\left|S^{(ref)} - S^{(comp)}\right|}.$$

Section 7 — Ambient happiness for keywords and short phrases

This section asks: instead of “How happy is Twitter today?”, ask “What is the emotional tone of tweets about X, over time?”

7.1 Definition: what is “ambient happiness”?

Take a keyword (like “BP” or “Tiger Woods”).

1. Collect all tweets containing that keyword → that’s your text T.
2. Compute the happiness of that text.
3. Optionally remove the keyword itself from the text before computing (so the keyword’s own rating doesn’t dominate).

They define “ambient happiness” as a *relative* measure: how happy tweets containing the term are compared to a baseline. The goal is to measure the emotional “surround” of the topic.

Table 2 (and related tables): reading the keyword list as a cultural selection

The authors choose a list of 100 “text elements” (single words or short phrases) and compute their ambient happiness. This is not a randomly sampled list; it’s a *curated* list meant to be interpretable (names, emotions, institutions, and a few pop-culture topics).

How to read the table:

- Treat it as a **ranking produced by the instrument** (“tweets containing X tend to be happier/sadder than baseline”).
- Then immediately ask: **what kinds of things end up at the top or bottom, and why?**

Patterns the authors point out (and you can check as you read):

- Very happy elements include affectionate or positive social tokens (some emoticons, “happy,” “love,” etc.).
- Very unhappy elements include terms linked to conflict, illness, and trauma.
- Political and institutional terms often sit below baseline (not always “unhappy,” but relatively less happy).

The table is also a snapshot of what “counted” as salient topics on Twitter in this period, and what the authors thought would be meaningful categories.

Figure 14: ambient happiness *and* how often a term appears

Figure 14 is easy to misread if you only look at the happiness curve.

It has two key ideas:

1. **Ambient happiness time series** (panel A): “When people tweet using this term, what is the average emotional tone of the surrounding words?”
2. **Occurrence frequency** (panel B): “How common is this term on Twitter over time?”

How to interpret the two panels together:

- If the **frequency spikes** at the same time that **ambient happiness changes**, that often signals a real-world event driving both attention and tone.
- A term can become *more common* without becoming *more positive* (or vice versa).
- Frequency is a reminder that *measurement reliability changes with volume*: if a term is rare in a month, its ambient happiness can be noisier.

As a reader: ask what kinds of events would change **who is tweeting** about a topic, not just what they feel about it (e.g., scandals, disasters, elections).

Figure 15: happiness vs lexical diversity are mostly uncorrelated (for these 100 terms)

Figure 15 compares two rankings for the same 100 text elements:

- rank by **ambient happiness**
- rank by **Simpson lexical size** (effective vocabulary diversity in the surrounding tweets)

The authors’ claim: there is little correlation. In plain language:

> Some topics generate diverse vocabulary but not unusually happy language; other topics generate happier language but not unusually diverse vocabulary.

Why this matters:

- It supports the idea that “information/diversity” and “happiness” are **different dimensions** of language.
- It’s also a warning: if you see a happiness change, don’t assume it is “just” topic diversity (and vice versa).

Figures 16–17: worked examples of keyword time series

These figures show time series for specific topics and then word shifts that explain spikes.

The reading lesson: the same instrument can be applied to subsets, but the meaning of the signal changes—because you are now measuring a *topic-shaped slice* of the platform.

Section 8 — Concluding remarks

- Big-data text streams make description and pattern-finding a primary task.
- Instruments like the hedonometer can complement surveys.
- Researchers must stay aware of manipulation and representativeness issues.