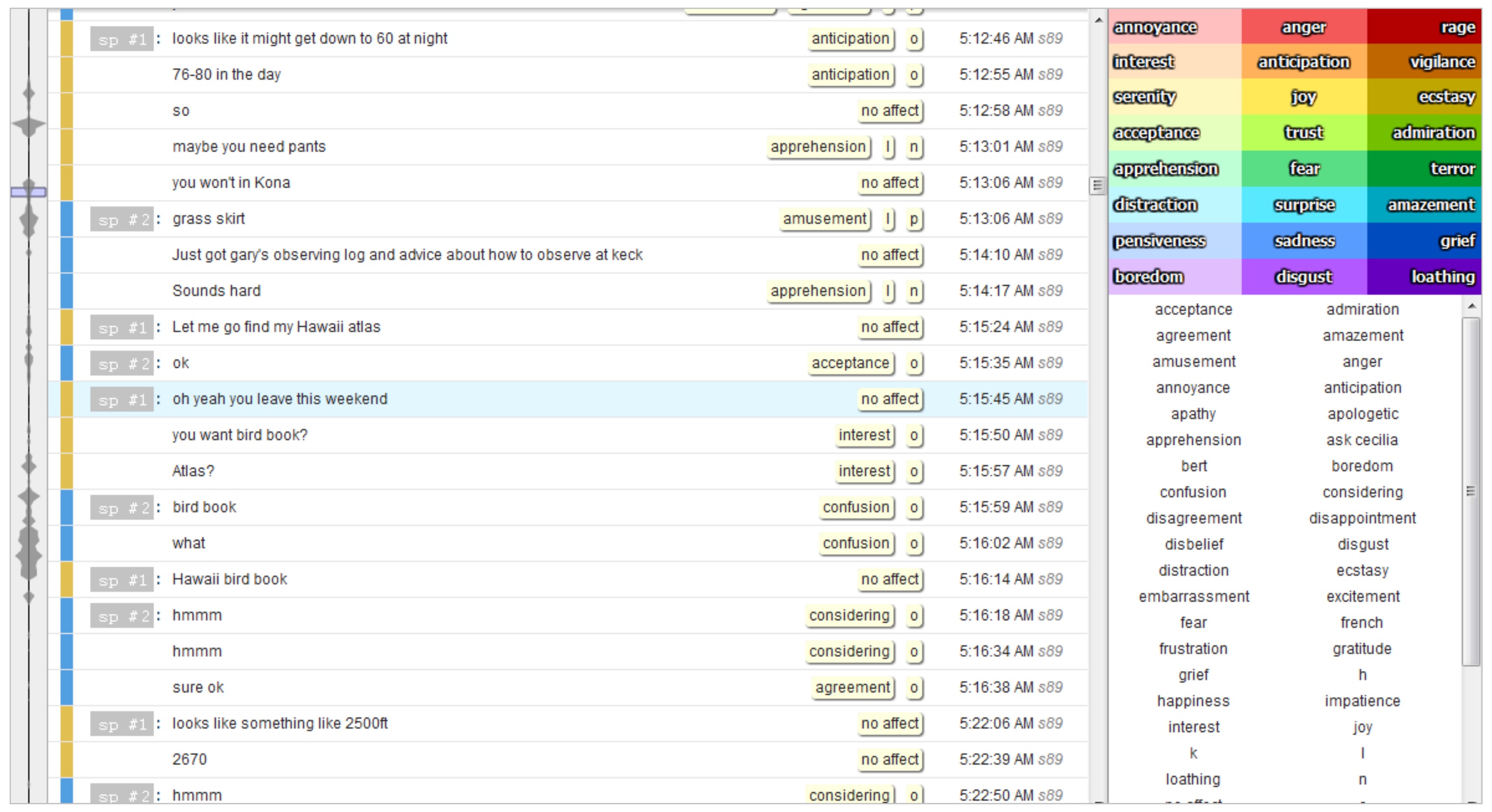
Text Classification for Qualitative Analysis at Scale

Michael Brooks, Katie Kuksenok, Megan Kelly Torkildson, John J. Robinson, Ray Hong, Taylor J. Scott, Daniel Perry, Cecilia Aragon
University of Washington
{mjbrooks, kuksenok, mtorkild, soco, rayhong, omni, dbperry, aragon}@uw.edu

Human Centered Text Analysis

How can computational analysis tools like **machine learning** and **visualization** support **qualitative analysis** of large text datasets?

The wealth of data generated online through social media, forums, and email is creating new areas and paradigms for the social sciences. However, obtaining and working with datasets at this scale requires particular skills and resources. We need **better tools for broader participation** across research communities. We are using a human-centered approach to design and build software supporting large-scale qualitative research.



sp #1: looks like it might get down to 60 at night — anticipation | o — 5:12:46 AM s89
76-80 in the day — anticipation | o — 5:12:55 AM s89
so — no affect — 5:12:58 AM s89
maybe you need pants — apprehension | l | n — 5:13:01 AM s89
you won't in Kona — no affect — 5:13:06 AM s89
sp #2: grass skirt — amusement | l | p — 5:13:06 AM s89
Just got gary's observing log and advice about how to observe at keck — no affect — 5:14:10 AM s89
Sounds hard — apprehension | l | n — 5:14:17 AM s89
sp #1: Let me go find my Hawaii atlas — no affect — 5:15:24 AM s89
sp #2: ok — acceptance | o — 5:15:35 AM s89
sp #1: oh yeah you leave this weekend — no affect — 5:15:45 AM s89
you want bird book? — interest | o — 5:15:50 AM s89
Atlas? — interest | o — 5:15:57 AM s89
sp #2: bird book — confusion | o — 5:15:59 AM s89
what — confusion | o — 5:16:02 AM s89
sp #1: Hawaii bird book — no affect — 5:16:14 AM s89
sp #2: hmmm — considering | o — 5:16:18 AM s89
hmmm — considering | o — 5:16:34 AM s89
sure ok — agreement | o — 5:16:38 AM s89
sp #1: looks like something like 2500ft — no affect — 5:22:06 AM s89
2670 — no affect — 5:22:39 AM s89
sp #2: hmmm — considering | o — 5:22:50 AM s89

annoyance	anger	rage
interest	anticipation	vigilance
serenity	joy	ecstasy
acceptance	trust	admiration
apprehension	fear	terror
distraction	surprise	amazement
pensiveness	sadness	grief
boredom	disgust	loathing

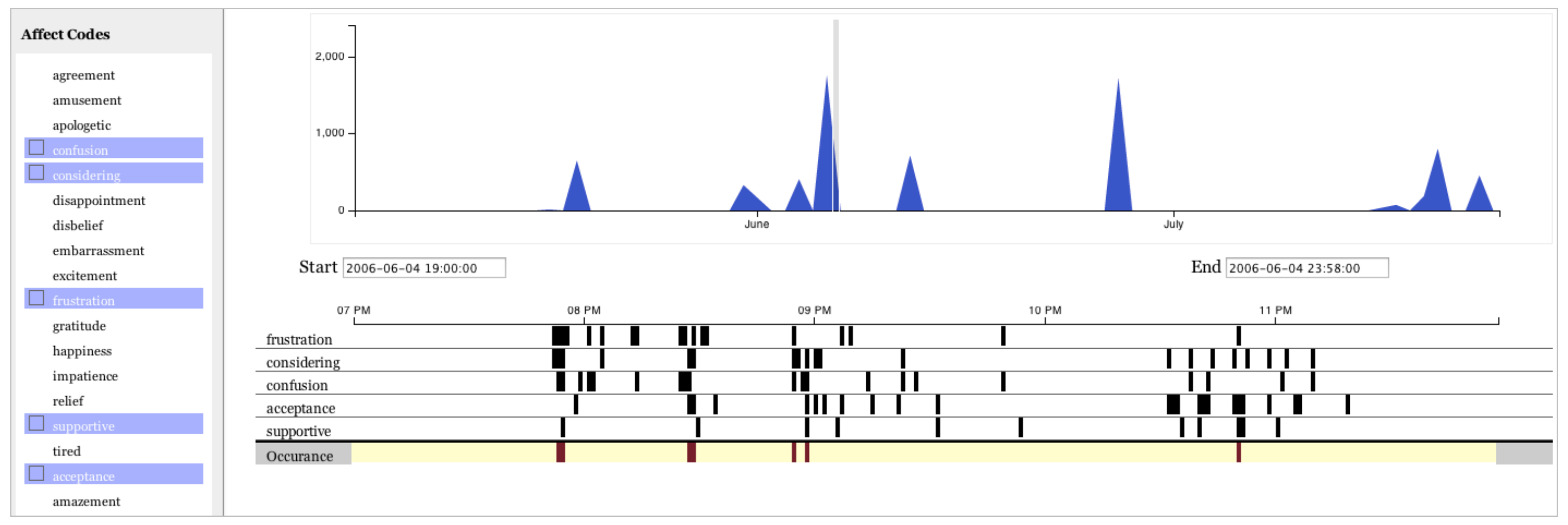
A tool we created for coding chat messages more efficiently. Visualizing message frequency over time shows the pace of conversations. Color-coding speakers reveals the conversation thread. The grid of codes helps users remain cognizant of the range of available affect categories.

Qualitative Data Analysis

- Many approaches to qualitative research exist.
- Place emphasis on **context, meaning, and interpretation**.
- Explicitly acknowledge and reflect on the **role of the researcher**.

Researchers often combine qualitative and quantitative approaches, but qualitative methods are built on different foundations from quantitative methods. This has implications for design.

Qualitative analysis relies on reading, coding, and summarizing text, **time-consuming** but critical meaning-making activities. Qualitative researchers are often limited to small, focused datasets. We believe that machine learning and visualization can magnify or extend the power of qualitative analysis.



A prototype visualization showing frequency of affect codes over time within our chat dataset. Visualizations such as this can show patterns in the categories which researchers could explore further.

Computational Social Science

- Social media and online communication tools produce vast amounts of text data: Twitter, email, online forums, and instant messaging.
- New research bridging social science and computer science is thriving on this public record of human social activity.
- How is activity online linked with offline phenomena such as community well-being and depression?
- How does information spread and propagate online?

While the research questions in this area are social, the methods are usually computational or statistical. We are building **analytics software grounded in qualitative methods**, to make large-scale social data more useful for research where a qualitative/interpretive approach is appropriate.

Text Classification

- *Text classification* has been applied to important problems including sentiment analysis, topic detection, and spam filtering.
- Decomposes raw text documents into numeric **features** such as the presence or absence of specific words and phrases.
- **Classifiers** can be trained to map these feature values onto a set of document classes.

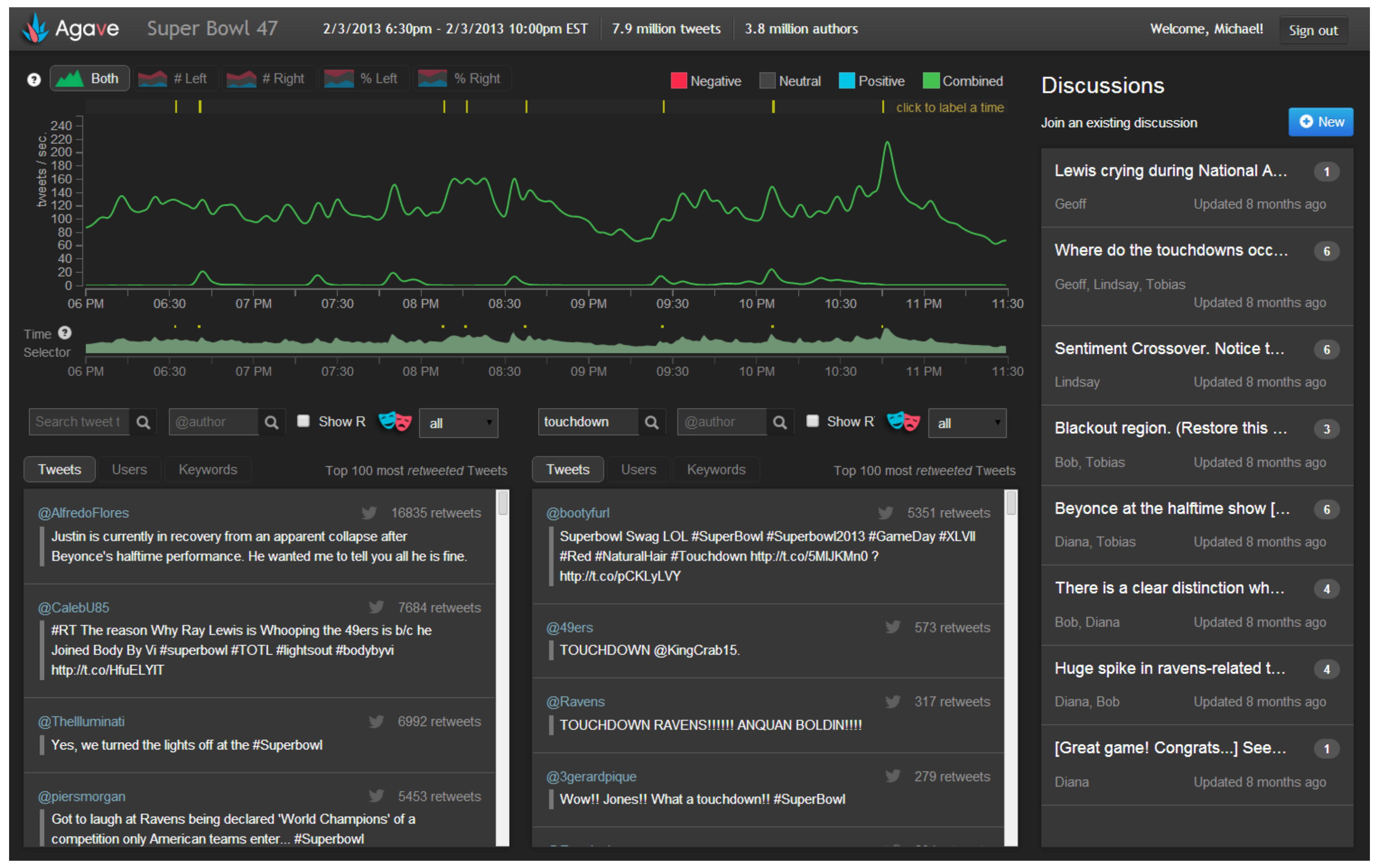
Trained classifiers are typically evaluated and optimized based on **quantitative metrics** such as accuracy, precision, or recall. This requires comparing the classifier's output to some pre-classified example documents, or ground **truth data**. It is not clear whether these approaches to evaluation are appropriate and sufficient when applied as part of qualitative analysis.

Visual Analytics

- Visual analytics is the use of interactive visualizations and computation for **analytical reasoning and data exploration**.
- Leverage the strengths of both humans and computers in a complementary fashion.
- Interactive visualizations could help researchers understand text-based datasets, feature spaces, and classifier performance.

Large, text communication datasets are typically complex, multifaceted, and time-varying. Interactive visualizations can be helpful for exploring the relationships between people, the mood and content of conversation, as well as dynamics over time.

From the initial open-coding phase of qualitative research, users may wish to begin experimenting with machine learning, whether to find new data to examine or to extend their coding work over more data. Designing for an interactive and exploratory style of use will enable non-experts to effectively apply machine learning to their own research problems. Guided by their intuition and domain knowledge, users will be able to rapidly create and evaluate classifiers with the support of visualizations and example results.



Agave | Super Bowl 47 | 2/3/2013 6:30pm - 2/3/2013 10:00pm EST | 7.9 million tweets | 3.8 million authors | Welcome, Michael! Sign out

@AlfredoFlores — 16835 retweets
Justin is currently in recovery from an apparent collapse after Beyonce's halftime performance. He wanted me to tell you all he is fine.

@CalebU85 — 7684 retweets
#RT The reason Why Ray Lewis is Whooping the 49ers is b/c he Joined Body By Vi #superbowl #TOTL #lightsout #bodybyvi http://t.co/HfuELYIT

@TheIlluminati — 6992 retweets
Yes, we turned the lights off at the #Superbowl

@piersmorgan — 5453 retweets
Got to laugh at Ravens being declared 'World Champions' of a competition only American teams enter... #Superbowl

@bootyfurl — 5351 retweets
Superbowl Swag LOL #SuperBowl #Superbowl2013 #GameDay #XLVII #Red #NaturalHair #Touchdown http://t.co/5MlJKMn0 ? http://t.co/pCKLyLVY

@49ers — 573 retweets
TOUCHDOWN @KingCrab15.

@Ravens — 317 retweets
TOUCHDOWN RAVENS!!!!!! ANQUAN BOLDIN!!!!

@3gerardpique — 279 retweets
Wow!! Jones!! What a touchdown!! #SuperBowl

Discussions
Join an existing discussion

Lewis crying during National A... — 1 — Geoff — Updated 8 months ago
Where do the touchdowns occ... — 6 — Geoff, Lindsay, Tobias — Updated 8 months ago
Sentiment Crossover. Notice t... — 6 — Lindsay — Updated 8 months ago
Blackout region. (Restore this ... — 3 — Bob, Tobias — Updated 8 months ago
Beyonce at the halftime show [... — 6 — Diana, Tobias — Updated 8 months ago
There is a clear distinction wh... — 4 — Bob, Diana — Updated 8 months ago
Huge spike in ravens-related t... — 4 — Diana, Bob — Updated 8 months ago
[Great game! Congrats...] See... — 1 — Diana — Updated 8 months ago

Agave, a system we built for collaborative exploration of large Twitter datasets. The screenshot shows a user inspecting tweets related to touchdowns in the 2013 Super Bowl, a 7 million tweet dataset. Tools that facilitate collaboration could be especially useful for research on large social data sets, which often involve multidisciplinary teams.

For more information, contact Michael Brooks at mjbrooks@uw.edu or Cecilia Aragon at aragon@uw.edu or visit http://depts.washington.edu/sccl

Scientific Collaboration & Creativity Lab
HCDE Human Centered Design & Engineering
dub
W UNIVERSITY of WASHINGTON
University of Washington Computer Science & Engineering