Comment Analysis

Nitesh Sekhar	(13CS10033)
Riya Bubna	(13CS10041)
Shrey Garg	(13CS10045)
Abhishek Niranjan	(13CS30003)
Mayank Roy	(13CS30021)

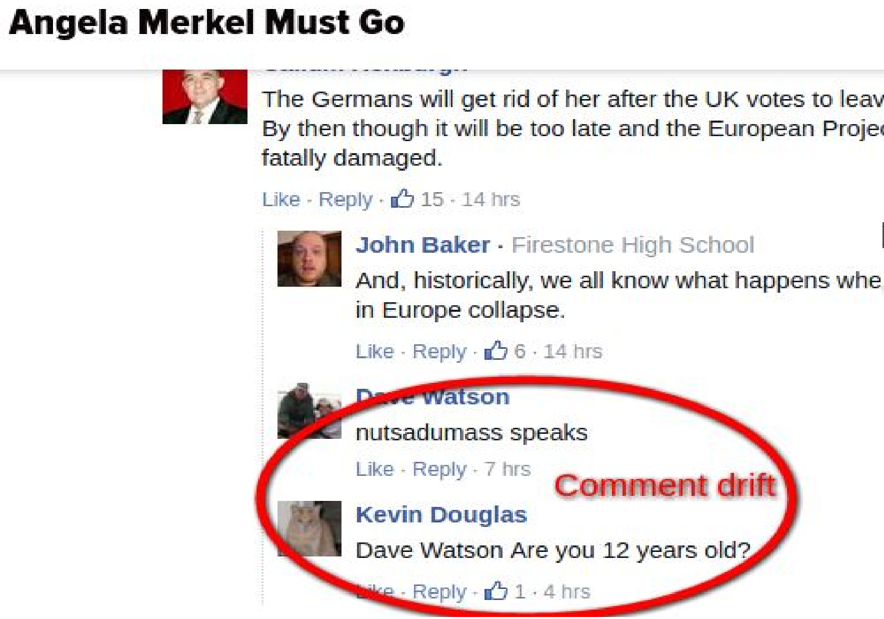
Motivation : Automoderator

- Automated Comment Section Moderation in any large data generating application or website.
- Prominent examples:
  - News websites attracting huge number of responses.
  - Online Stores using comment sections as a de-facto recommender system
  - Automatic moderation for AMA like activities or live telecasts
  - Comment moderation on sites like youtube.com and moderation being enabled on a case by case basis.

Motivation : Comment Drift

- In the previous idea, the final analysis would've been on a similarity measure between article-generated topics and comment-generated topics
- Now, in case of a highly viewed article, there can be specific comments because of which these similarity values can change over time because a specific central comment is driving the discussion in a new direction.
- This is the idea of comment topic drift
- Usecase:
  a. To identify key users for a site, since the users starting and driving these comments are the ones driving traffic to the site
  b. Using these specific users as markers for improvement in user retention

Examples



Angela Merkel Must Go

The Germans will get rid of her after the UK votes to leave the EU.
By then though it will be too late and the European Project will have been fatally damaged.

Like · Reply · 15 · 14 hrs

John Baker · Firestone High School

And, historically, we all know what happens whe
in Europe collapse.

Like · Reply · 6 · 14 hrs

Dave Watson

nutsadumass speaks

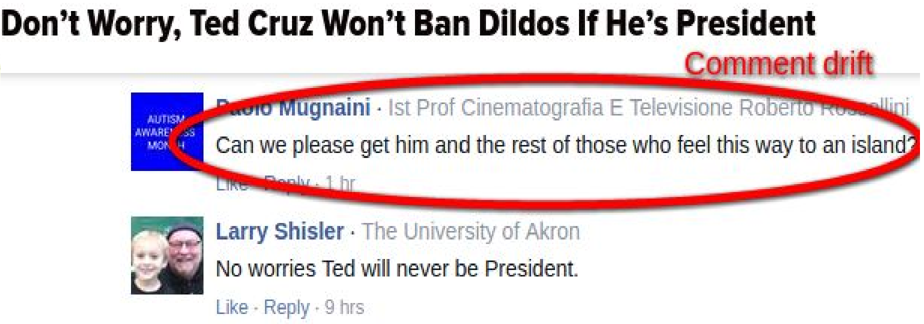
Like · Reply · 7 hrs

Comment drift

Kevin Douglas

Dave Watson Are you 12 years old?

Like · Reply · 1 · 4 hrs



Don't Worry, Ted Cruz Won't Ban Dildos If He's President

Comment drift

Paolo Mugnaini · Ist Prof Cinematografia E Televisione Roberto Rossellini

Can we please get him and the rest of those who feel this way to an island?

Like · Reply · 1 hr

Larry Shisler · The University of Akron

No worries Ted will never be President.

Like · Reply · 9 hrs

Working and Testing : Data Crawling

- The data has been crawled from websites like The Guardian, The Huffington Post, The Atlantic, Fox News, and others.
- This crawler was run on the server to get all comments by clicking on the relevant buttons for comments, load more comments etc.
- We used selenium and BeautifulSoup to crawl all the data.

S.No	Sites	Categories	Article_Num	Avg_Comment_n	Sub_comments	User_data	Tested on
1	The Guardian	Politics, Sports, Entertainment	1042	50	Yes	Yes	Code, tested
2	Huffington Post	Politics, Science, Entertainment	30	200	Yes	Yes	Code, tested
3	The Atlantic	Politics	20	400	Yes	Yes	Code, Tested
4	Fox News	Politics, Entertainment	172	50	No	No	Code, not tested
5	NDTV	Politics,	40	100	No	No	Code, not tested

Working and Testing : Natural Language Processing

The raw dataset which was crawled from the websites needed a lot of processing:

- Cleaning to convert text to a processable format
- Tokenization
- Stemming
- Removing Stopwords
- Sentence segmentation to generate documents for the LDA

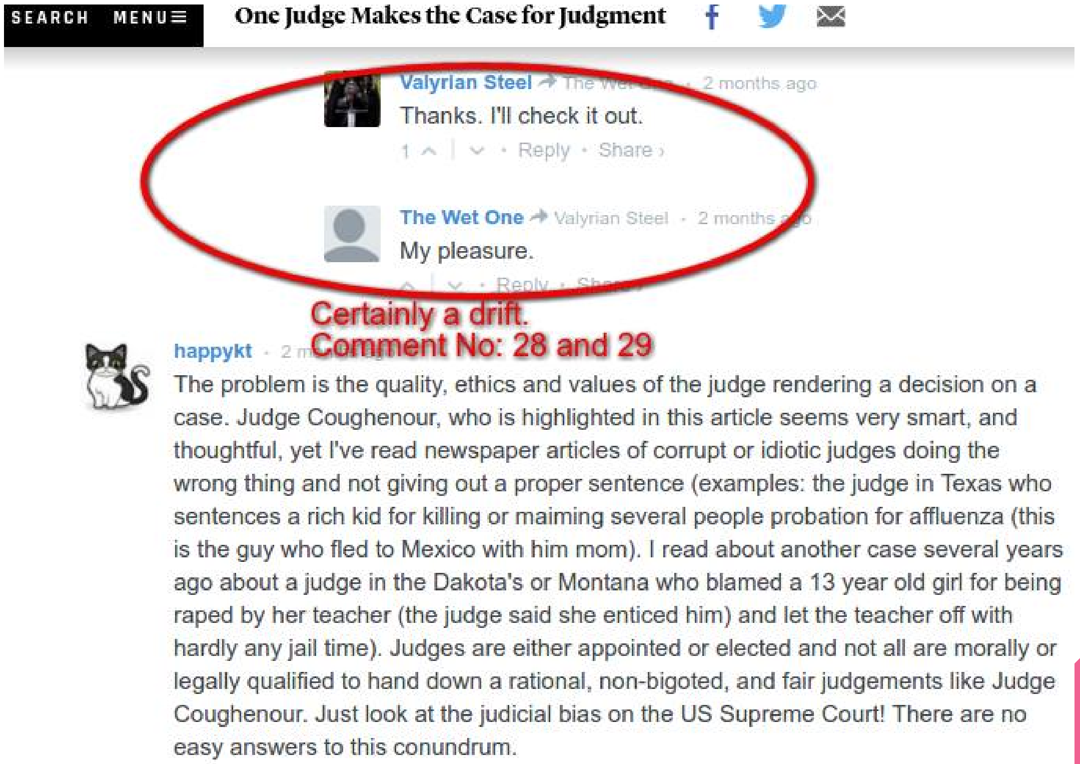
OUTPUT FORMAT:

```
<headline>
        <title of the article>
<article>
        <sentence 1>
        <sentence 2>...
<comments>
        <Username 1>
        <comment 1>
        <Username 2>
        <comment 2>...
```

Working and Testing : Centrality Analysis

- **Latent Dirichlet Allocation Model**
  - Used to find the topics associated with each article where each document is either 1 or 2 sentences of the article.
  - LDA was also used to find which document or sentence corresponded to which topic
- **Similarity Measure Used - Tf Idf**
  - Used to compute similarity between topics and comments
  - Used to calculate these heuristics:
    - Comment to topic mapping
    - Comment to sentence mapping
    - User to topic mapping
    - User to sentence mapping

Workflow: Article



SEARCH MENU≡ **One Judge Makes the Case for Judgment**

Valyrian Steel → The Wet One · 2 months ago

Thanks. I'll check it out.

1 ^ | ˅ · Reply · Share ›

The Wet One → Valyrian Steel · 2 months ago

My pleasure.

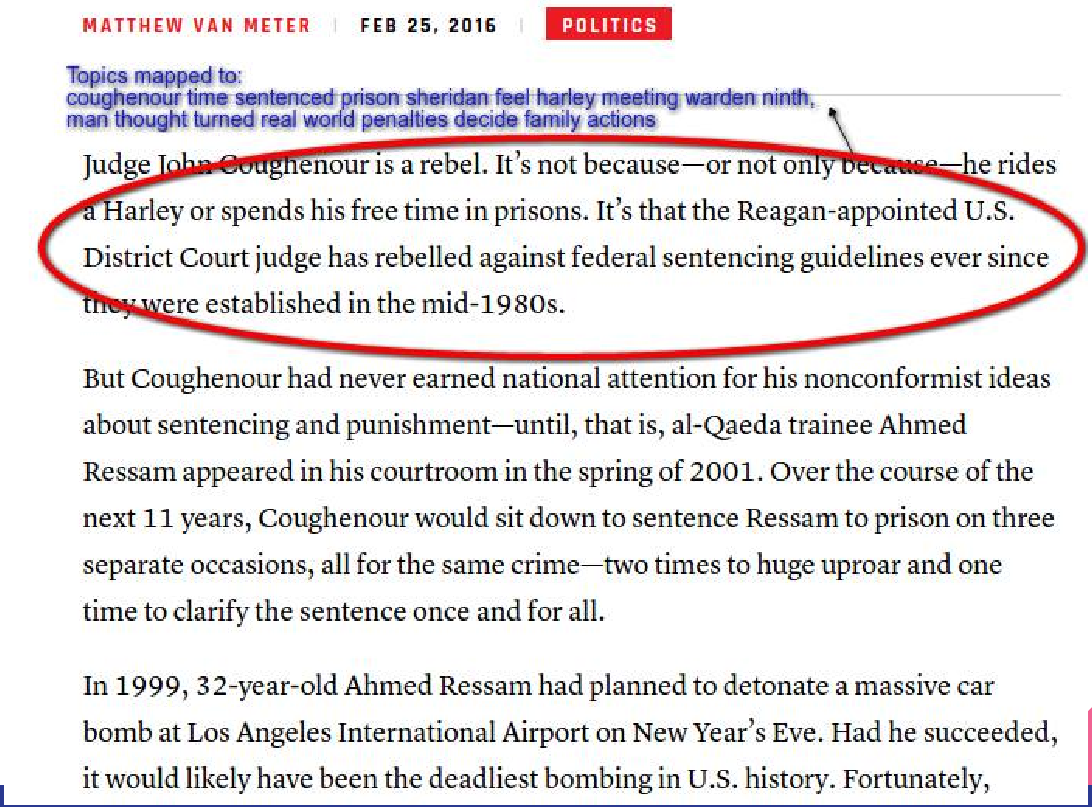
^ | ˅ · Reply · Share ›

Certainly a drift.
Comment No: 28 and 29

happykt · 2 months ago

The problem is the quality, ethics and values of the judge rendering a decision on a case. Judge Coughenour, who is highlighted in this article seems very smart, and thoughtful, yet I've read newspaper articles of corrupt or idiotic judges doing the wrong thing and not giving out a proper sentence (examples: the judge in Texas who sentences a rich kid for killing or maiming several people probation for affluenza (this is the guy who fled to Mexico with him mom). I read about another case several years ago about a judge in the Dakota's or Montana who blamed a 13 year old girl for being raped by her teacher (the judge said she enticed him) and let the teacher off with hardly any jail time). Judges are either appointed or elected and not all are morally or legally qualified to hand down a rational, non-bigoted, and fair judgements like Judge Coughenour. Just look at the judicial bias on the US Supreme Court! There are no easy answers to this conundrum.

Workflow: Topic Generation



MATTHEW VAN METER | FEB 25, 2016 | POLITICS

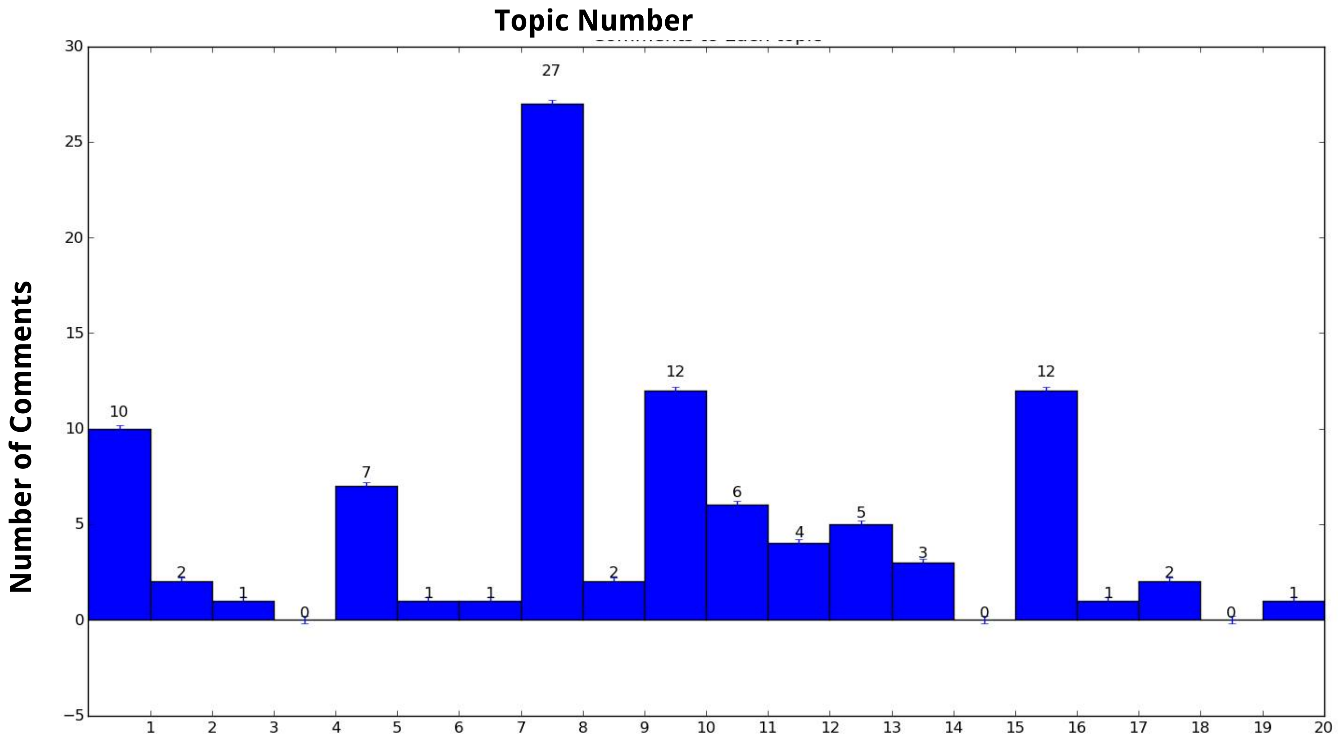
Topics mapped to:
coughenour time sentenced prison sheridan feel harley meeting warden ninth,
man thought turned real world penalties decide family actions

Judge John Coughenour is a rebel. It's not because—or not only because—he rides a Harley or spends his free time in prisons. It's that the Reagan-appointed U.S. District Court judge has rebelled against federal sentencing guidelines ever since they were established in the mid-1980s.

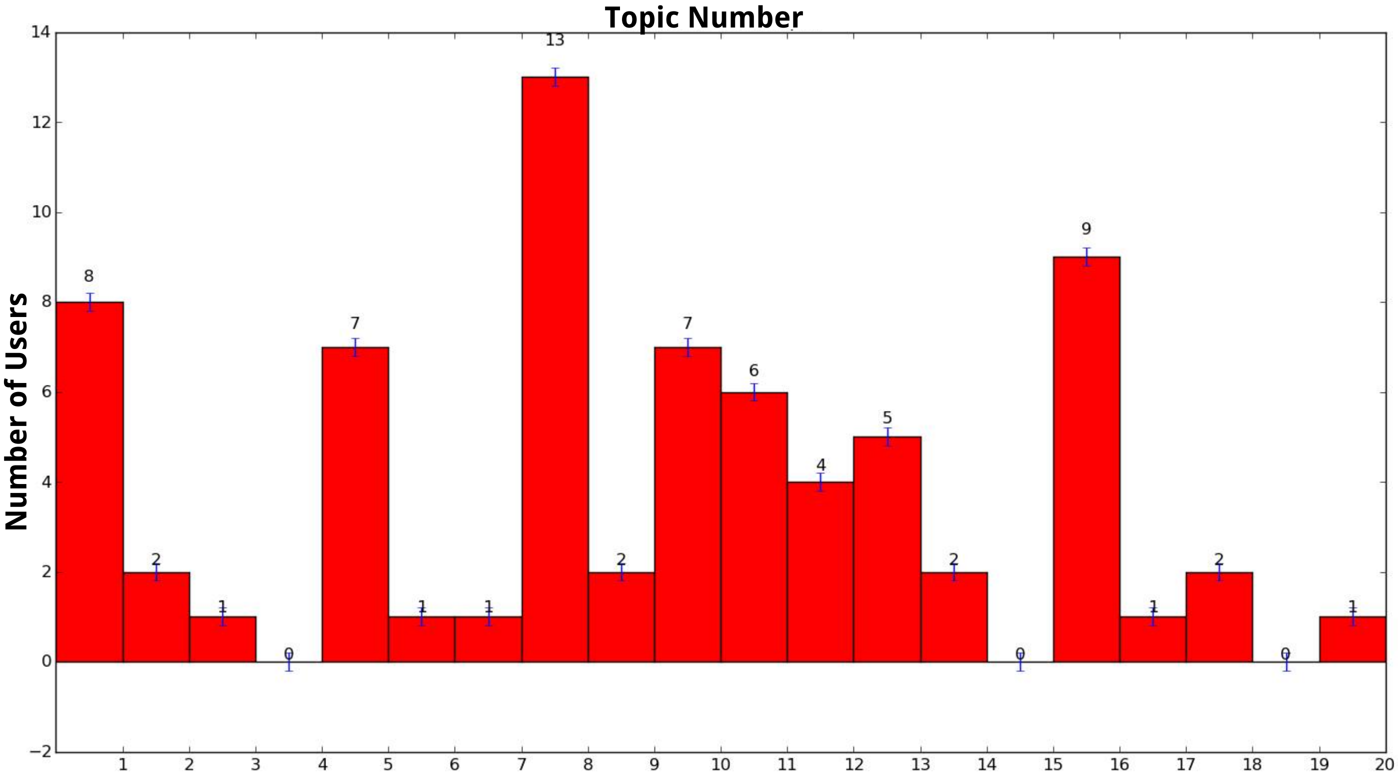
But Coughenour had never earned national attention for his nonconformist ideas about sentencing and punishment—until, that is, al-Qaeda trainee Ahmed Ressam appeared in his courtroom in the spring of 2001. Over the course of the next 11 years, Coughenour would sit down to sentence Ressam to prison on three separate occasions, all for the same crime—two times to huge uproar and one time to clarify the sentence once and for all.

In 1999, 32-year-old Ahmed Ressam had planned to detonate a massive car bomb at Los Angeles International Airport on New Year's Eve. Had he succeeded, it would likely have been the deadliest bombing in U.S. history. Fortunately,

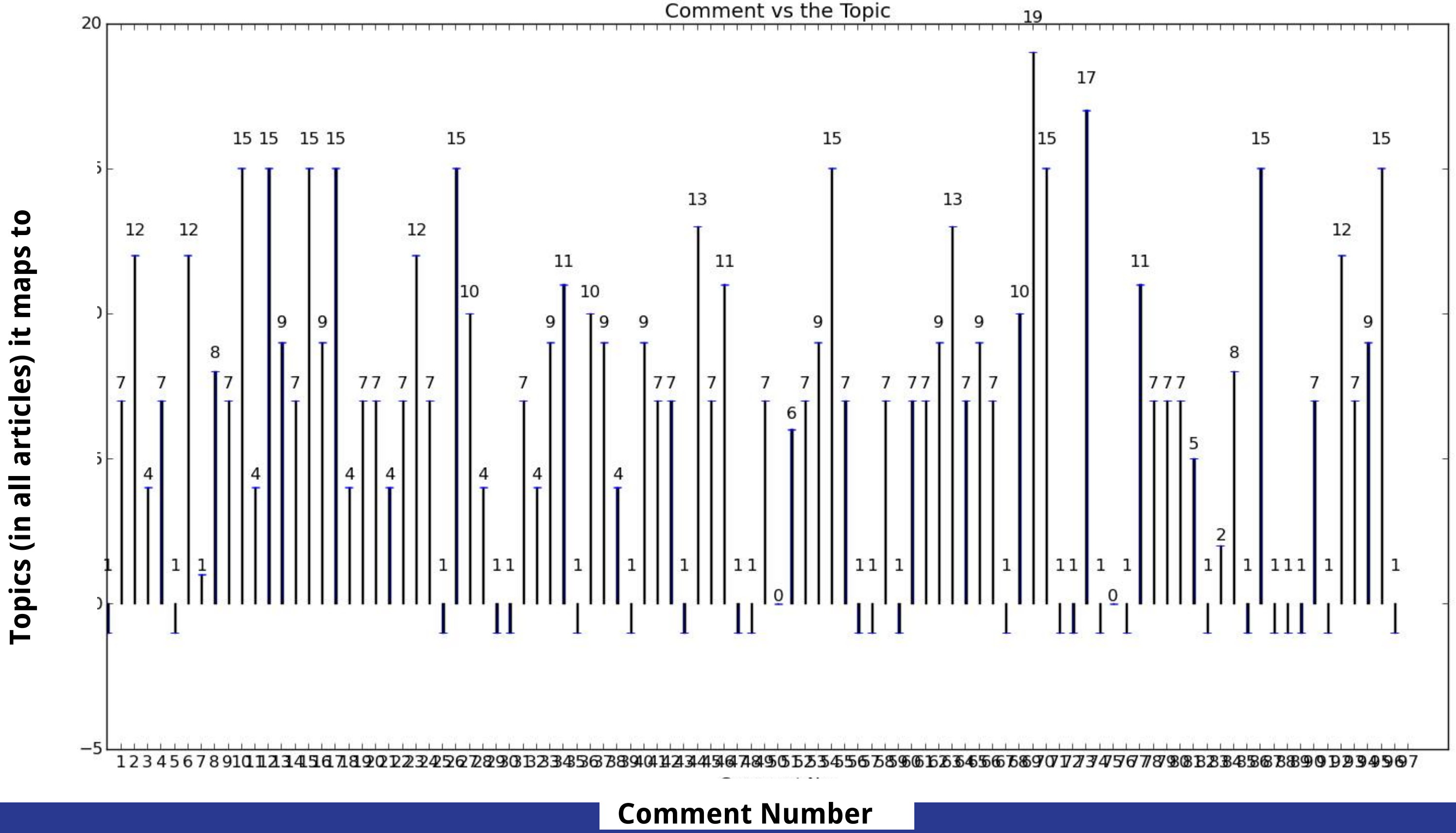
Workflow: Topic to comment mapping



Workflow: Topic to user mapping



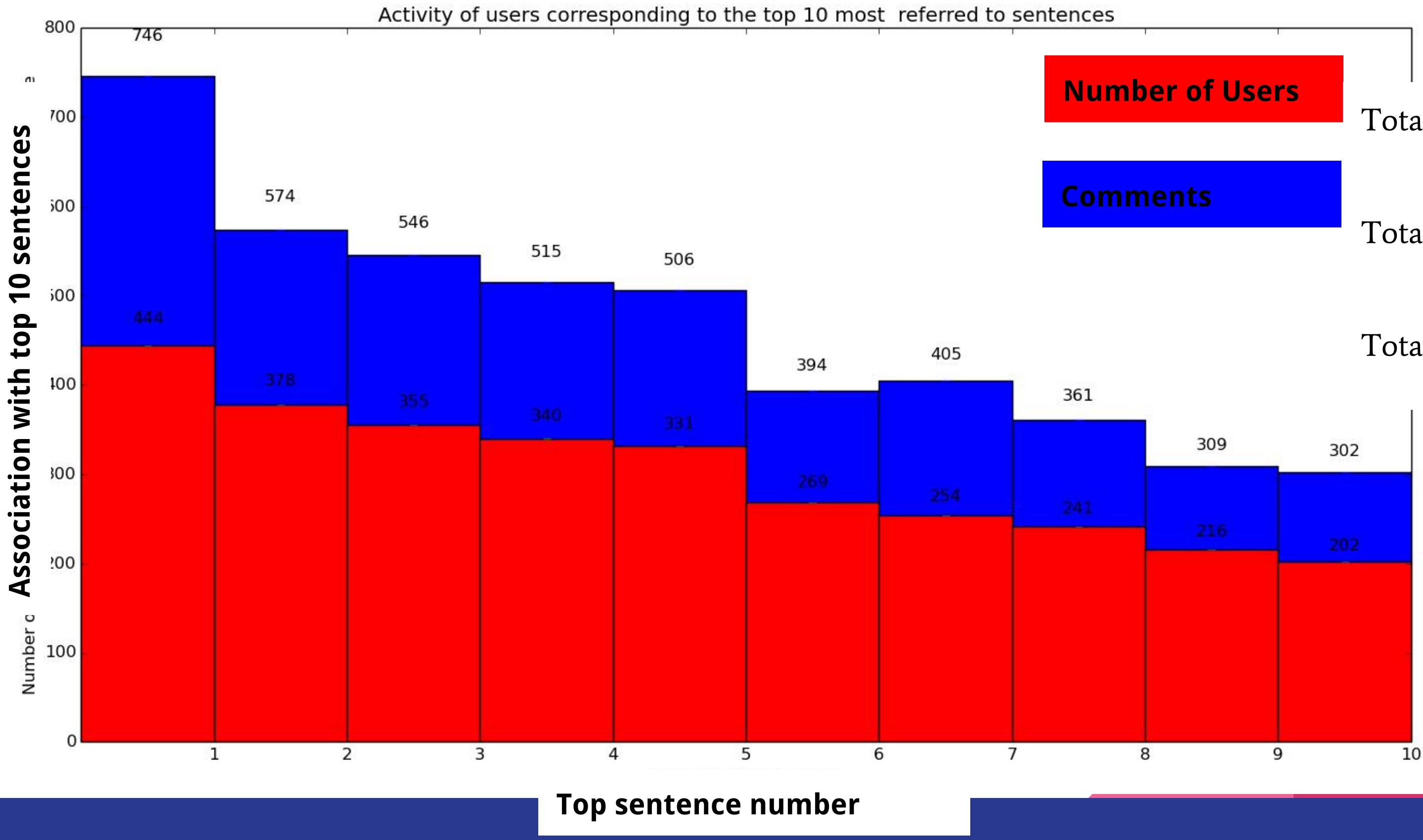
Workflow: Comment to Topic Mapping



Workflow: Analysis of the output

- Certain topics are most debated/talked about in the comment section.
- The number of new users engaging in any topic is highest for the topics which already have many comments related to it showing **herd mentality**.
- There are two types of comment drifts:
    - Intra-document drift
    - Comments totally unrelated to the article.

Sentence to Comment and User Mapping

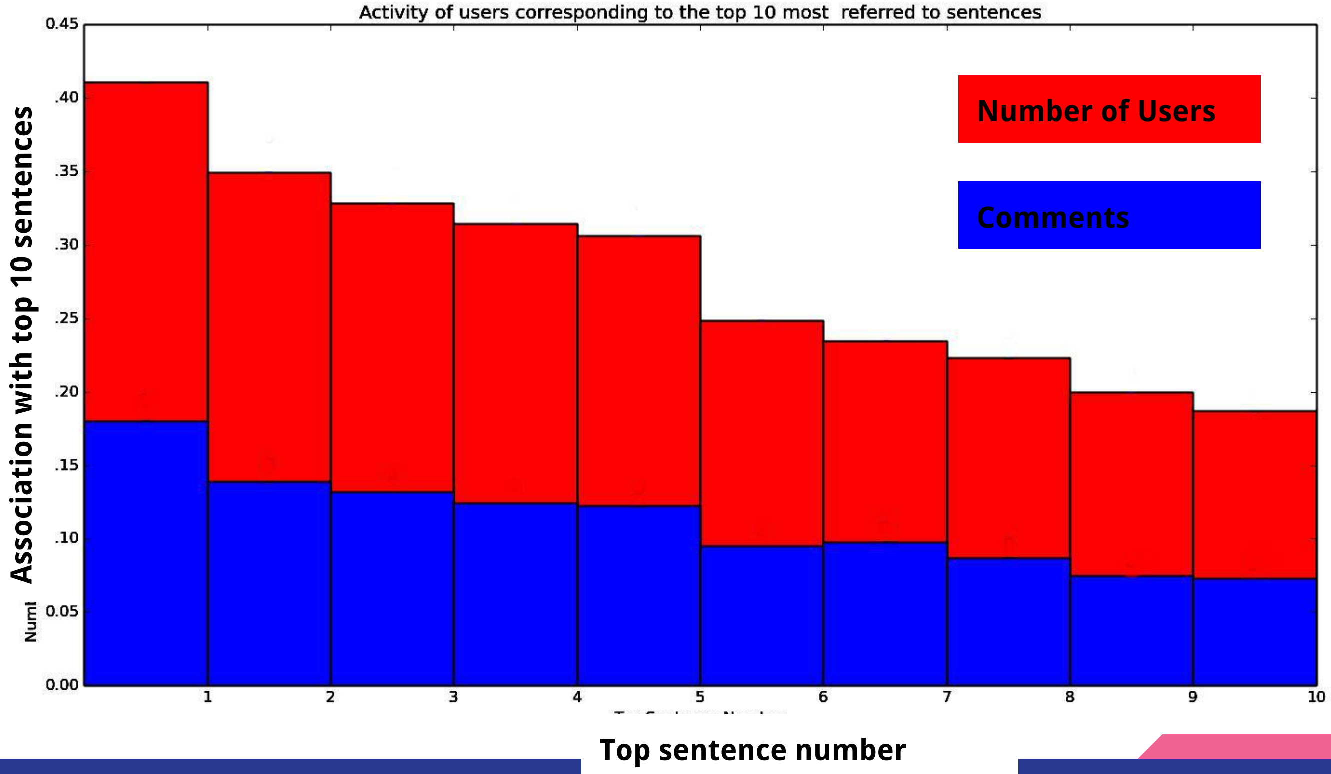


Total Articles : 20

Total Comments : 4097

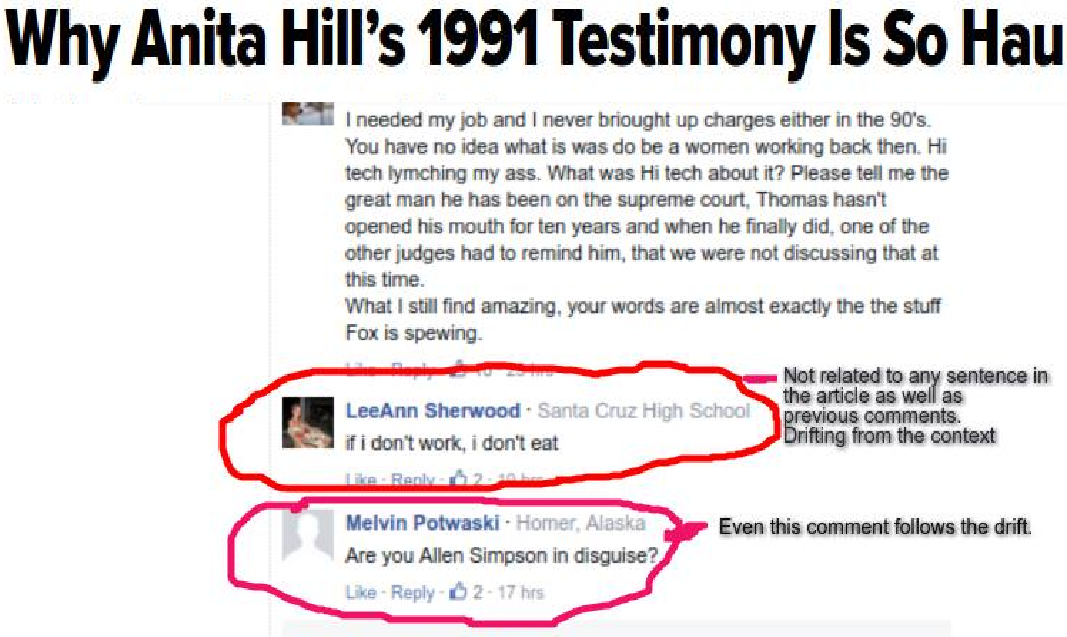
Total Users : 1182

Analyzing the most engaging sentences



Results

- A clip showing how the comments drift from the context of the news articles. This particular article had 29 comments out of which, 5 were unrelated comments.



Why Anita Hill's 1991 Testimony Is So Haunting Today

I needed my job and I never briought up charges either in the 90's. You have no idea what is was do be a women working back then. Hi tech lymching my ass. What was Hi tech about it? Please tell me the great man he has been on the supreme court, Thomas hasn't opened his mouth for ten years and when he finally did, one of the other judges had to remind him, that we were not discussing that at this time.
What I still find amazing, your words are almost exactly the the stuff Fox is spewing.

Like · Reply · 10 · 23 hrs

LeeAnn Sherwood · Santa Cruz High School
if i don't work, i don't eat
Like · Reply · 2 · 19 hrs

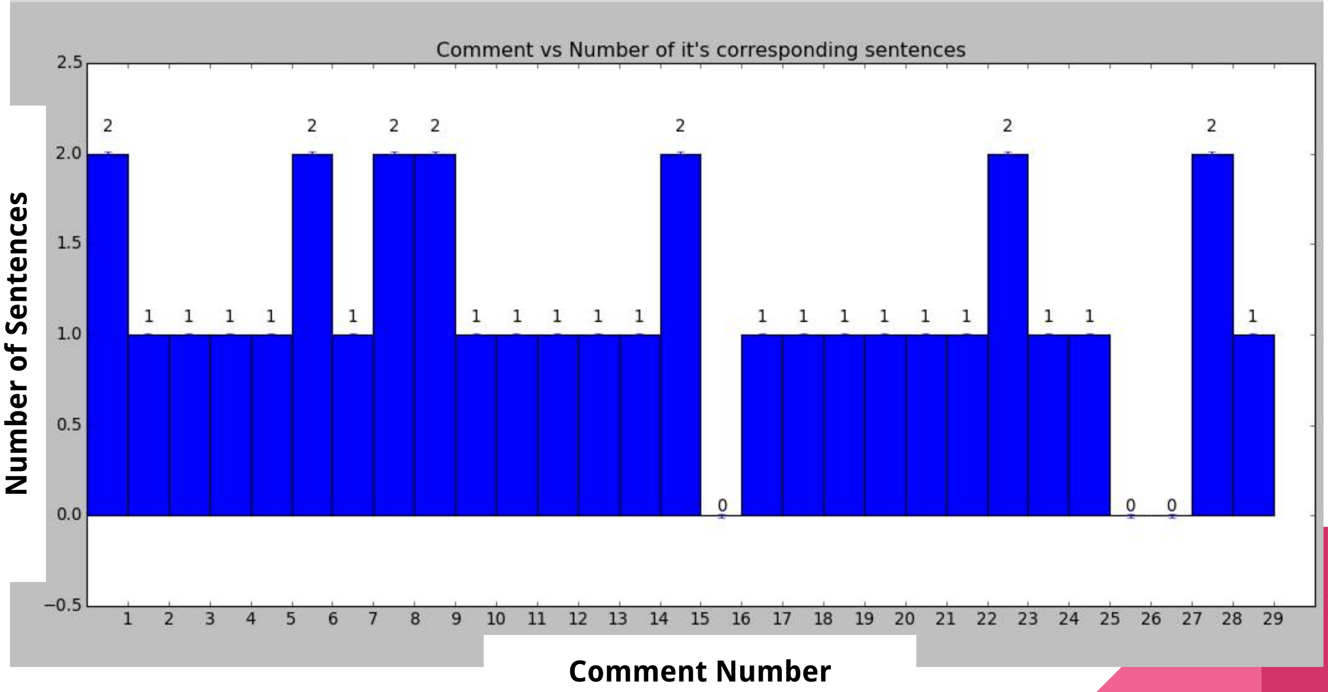
Not related to any sentence in the article as well as previous comments. Drifting from the context

Melvin Potwaski · Homer, Alaska
Are you Allen Simpson in disguise?
Like · Reply · 2 · 17 hrs

Even this comment follows the drift.

Results

Out of those 5 unrelated comments, 3 were detected by our code.



Conclusions and Future Plans

- Two types of drifts are identified within an article:
  - Inter-Topics Drift
  - Out of Context(Document) Drift
- The highest referred to sentences, have a significantly greater fraction of Users and Comments associated with it.
- Only primary level of thresholding has been applied to filter the relevant topics.
- Plan to apply algorithm to find Hub and Authority centrality of comments for the next level of filtering of negative positives.
- The idea is to supervise a model which automatically learns about the first comment that will cause drift to occur.

THANK YOU!!