Twitter, Slashdot, Imitation, Obfuscation, (and a little bit of) Regulation.

Michael Brennan
Drexel University
Privacy, Security and Automation Lab
Overview

• Privacy, Security & Automation.
• Social Media and Artificial Intelligence
  • Using AI to Improve Discourse (Slashdot)
  • Machine Learning to Improve Classification (Twitter)
• Privacy and Stylometry
  • Attacking Authorship Recognition to Protect Privacy
• Geeks in Washington
  • A Technologist at the Federal Trade Commission
Improving Online Discourse

- Crowdsourced filtering systems have been developed to manage information overload and improve discourse in online communities.
  - Goal: Good content will rise to prominence, bad content will fade into obscurity.
- Our research predicts ratings of discussion comments with the ultimate goal of augmenting such systems.
- We achieve 82% accuracy at extracting the best comments on Slashdot.org.
Questioning Crowdsourced Filtering

• The “Wisdom of Crowds” or a “censoring mob?”
• Which voices are amplified and which are silenced?
• How is influence and reputation built online?
• How easy is it to manipulate these systems?
Research Goals

• Replicate crowdsourced ratings in an online community.
  • Slashdot.org
  • Use machine learning to classify comments
• Lay the ground work for using agents to augment crowdsourced filtering.
  • Help identify content missed by the community
  • Important in communities that do not reach critical mass.
• Adjust for human weaknesses
  • Make better use of limited human oversight
Who here reads Slashdot?
Who here reads Slashdot?
The Slashdot Community

- Slashdot (slashdot.org) is a technology news site and online community.
- Readers submit articles which are reviewed by editors and the best ones are selected as news items.
- The community discusses each article through a comment system.
  - Moderation of discourse is facilitated by a robust rating system.
The Slashdot Rating System

• Subset of users are randomly given moderator points.
• Comments can be rated on a scale of -1 to 5.
• Comments can also be assigned positive descriptions such as “Insightful” or “Informative” and negative ones such as “Offtopic” or “Flamebait.”
• Many other nuances:
  • Anonymity penalty
  • Meta-moderation
  • Friends and enemies
  • Posting guidelines

“I USED TO FIND SLASHDOT DELIGHTFUL, BUT MY FEELINGS OF LATE ARE MORE SPITEFUL; MY COMMENTS SARCASTIC THE ICONOCLASTIC KEEP MODDING TO PLUS FIVE (INSIGHTFUL).”

“Fun game: try to post a YouTube comment so stupid that people realize you must be joking. (Hint: this is impossible)”
Rabbit Ears To Stage a Comeback Thanks To DTV

Posted by kdawson on Saturday February 14 2009, @05:55PM
from the standing-on-one-leg dept.

Jeffrey Breen writes

"Like Monty Python’s Killer Rabbit, cheap indoor antennas seem harmless to satellite and cable providers. But with the digital TV transition in the US, rabbit ears can suddenly provide digital-perfect pictures, many more channels, and even on-screen program guides. Already feeling pressure as suddenly budget-conscious consumers shed premium channels, providers must now get creative to protect their low-end as well."

business tv entertainment competition usa story

Not rabbit ears (Score:5, Informative)

by show me altoids (1183399) on Saturday February 14 2009, @05:57PM (#26858873)

Rabbit ears don't pick up UHF signals; they are for VHF which is going away. It's the "loop" part of current antennas which will receive UHF.

78 hidden comments
Features

• Context & Reputation
  • Time Difference, Sub Comments, Poster ID Number, Poster Acceptance Ratio.
  • All have been useful but some are only in evidence significantly after the comment has been posted.
  • Not necessarily scalable across domains.

• Linguistic
  • Sentiment, Swear Words, First Person Pronouns, Post Word Count.
  • Not as effective as contextual/reputation features.
  • Much easier to scale across domains.
Results: Extracting Best Comments

- Highlighting Top Comments: rating of 3 or higher.
  - 82% Accuracy
  - Default comment score is one, so a comment of 3 or higher can be considered to be very good.
  - Straightforward but important: highlighting the most relevant and worthwhile content.

- Predicting Bad (-1, 0), Neutral (1), Good (2-5) comments.
  - 76% Accuracy
  - Misclassifications skew appropriately – more difficult to distinguish between “neutral” and “good” than “bad” and “good.”

- Support Vector Machine Classifier.
Breakdown: Bad/Neutral/Good

<table>
<thead>
<tr>
<th># of Comments</th>
<th>Classified Bad</th>
<th>Classified Neutral</th>
<th>Classified Good</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bad [-1,0]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Neutral [1]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Good [2,3,4,5]</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Breakdown: Total comments for each classification type.
Misclassifications

• Good short comments often misclassified.
• Anonymous comments easier to classify: more likely to be rated negatively.
• Misclassified good comments are often from authors with little or negative reputation.
• Funny comments difficult to classify.

• “Blinking dead people! HOLY BLINKING DEADMAN!!! I for one welcome our blinking undead overlords.” (2, Funny)
• “Slashdotters? Oh please. It's the same as Diggers and Redditors. Nerds. Internet nerds. We're all atheist. We're all mostly libertarian. On and on.” (0, Offtopic)
What Worked? What Can We Learn?

• Contextual and Author Reputation features worked best.
  • Demonstrate how community structure can make a hard classification problem tractable.
  • But not completely satisfying as linguistic features offer greater scalability across domains.

• Despite difficulties of linguistic features, some were useful.
  • Second person pronouns more likely to indicate a bad post, first person pronouns indicate a positive post. Respect for ownership?
Open Questions

• How do these results translate to other communities?
• How can we combine human and machine computation to reduce information overload?
• How do online communities develop reputations, norms, and tastes?
• NLP on semi-structured data – is there a ceiling to determining if comments are good without understanding them?
• Do the features we selected really represent what a community is looking for?
• How will a community respond to an augmented system? Create better content? Hack “better” posts?
Improving Social Media

• Twitter & Trending Topics.
Improving Social Media

• Effectively understanding and analyzing social media for highly discussed topics benefits everyone.

• Recent events and new research show the increasingly important role of social media (and twitter specifically).
  • Haiti Earthquake
  • Uprising in Egypt
  • Human Rights in China
  • Ushahidi.com – Crowdsourcing Crisis Information

• Plus the traditional benefactors of such technology.
  • Advertisers (duh)
  • The users!
Twitter’s Trend Classification Problem

- Trends on Twitter are heavily discussed topics.
- Tweets are identified as being part of a trend if they contain the trend keyword ("Egypt" or "iPhone")
- Problem: Keywords are not enough to identify all tweets that are part of a trend.
- Solution: Use machine learning to identify which trend a tweet belongs to without the use of the trend keyword.
Trending Topics

• Top 10 terms on Twitter (common terms removed – “coffee”)
• Tweets with *term* or *#term* in them are part of the trend.
• Top worldwide trends on the Twitter homepage 5.02.10:
  • #willgetyoukilled
  • #nowplaying
  • #ICouldNeverDate
  • Brandy
  • #familyguy
  • Family Guy
  • Justin Bieber
  • #YouWillNever
  • #BieberRuinedTwitter
  • Boondocks

Real-time results for boondocks OR #Boondocks

Inspire1906 Counting down the minutes til the Boondocks premiere....
less than 20 seconds ago from TweetCaster

Sed_So @BruceDwayne send the links 4 the boondocks stream sites
1 minute ago from web

CWise818 Ok Y'all Gotta Catch Y'all Later got about 15mins till the Boondocks Season 3 comes on!!
2 minutes ago from Tweetie

K3WLK1D bam bam bam bam bam CHIL boondocks.
s000000 hype d up right now.
2 minutes ago from web
Trending Topics

• Catch that?
  • #familyguy, Family Guy
  • Justin Bieber, #BieberRuinedTwitter

• One distinction is correct, the other is not.
  • Family Guy and #familyguy are likely both referring to the 150th episode of the show, which aired that night.
  • Justin Bieber and #BieberRuinedTwitter are opposing views of the same topic, so maybe they should be separate!

• What about tweets without any trending term?
  • #Egypt might be popular, but a tweet saying “Mubarak thugs are violently suppressing protesters in Tahrir Square” would be missed.

• Bottom line: keywords aren’t enough.
Research Goals

• Can we get a reasonable level of accuracy by classifying tweets as being a part of a trend WITHOUT knowing the trend keyword?
  • What about with different numbers of trends? And complicated tweets?
• Evaluate the ability for straightforward machine learning to improve understanding and classification in social media.
Methodology

• Collected 43K tweets from 30K users for 30 trends over 3 days.
  • No “whitelist” status.
  • Trending topic keywords removed from all tweets.
  • “Clean” data set of 24K tweets only included those with more than 15 words/punctuation tokens and did not have multiple trend keywords.

• Used a modified Naïve Bayes Classifier
  • Transform Weight-Normalized Complement Naïve Bayes*
    • Mitigates Bayesian learning issues like skewed data bias and weight magnitude errors.

* Jason D. M. Rennie, Lawrence Shih, Jaime Teevan, and David R. Karger. “Tackling the Poor Assumptions of Naïve Bayes Text Classifiers.”
Correctly Identifying Trends

<table>
<thead>
<tr>
<th>Day</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wednesday</td>
<td>80%</td>
<td>80%</td>
</tr>
<tr>
<td>Thursday</td>
<td>80%</td>
<td>70%</td>
</tr>
<tr>
<td>Friday</td>
<td>80%</td>
<td>80%</td>
</tr>
</tbody>
</table>
Clean vs. Noisy Tweets

Precision
Recall

Original
Clean
10 Trends vs. 30 Trends

Precision

Recall

10 Trends

30 Trends
Misclassifications

• #EverLastingFriends argue.. maybe even fight dont talk for awhile but at the end of the day call n say “wat u doin”;
  • (#idontappreciate)
• Why is Zimbabwe trending?
  • (Almagro)
• Estriaaaa boa com esse hino!!! http://bit.ly/a9dnzB (Eclipse soundtrack)
  • (Eclipse Soundtrack)
• this is f’n outrageous! - http://cli.gs/h4UJGz
  #ifhiphopwashighschool Guachaca Zimbabwe
  #everlastingfriends Joran Data Plans Almagro
  • (#everlastingfriends)
Future Work in Social Media

• Future:
  • Greater analysis of misclassifications & features to avoid them.
  • Applying more modern machine learning techniques (SVMs)

• Conclusions
  • Demonstrates effectiveness of using machine learning to classify short messages.
  • Understanding and separating discourse in social media is important and machine learning can be used to make it more effective.
Privacy and Stylometry
What is Authorship Recognition?

- The basic question: “who wrote this document?”
- **Stylometry**: The study of attributing authorship to documents based only on the linguistic style they exhibit.
  - “Linguistic Style” Features: sentence length, word choices, syntactic structure, etc.
  - Handwriting, content-based features, and contextual features are not considered.
- In this presentation, stylometry and authorship recognition are used interchangeably.
Stylometry: the Threat to Privacy

- Good techniques for location privacy (Tor, Mixes, etc).
  - But it may be insufficient!
- Stylometry can identify authors based on their writing.
  - Can anonymous authors defend against this?
Supervised Stylometry

• Given a set of documents of known authorship, classify a document of unknown authorship.
  • Classifier trained on undisputed text.
• Scenario: Alice the Anonymous Blogger vs. Bob the Abusive Employer.
  • Alice blogs about abuses at Bob’s company.
    • Blog posted anonymously (Tor, pseudonym, etc).
  • Bob obtains 5000-10000 words of each employee’s writing.
    • Bob uses stylometry to identify Alice as the blogger.
Unsupervised Stylometry

• Given a set of documents of unknown authorship, cluster them into author groups.
  • No pre-existing author information.
  • “Similarity Detection”

• Scenario: Anonymous Forum vs. Oppressive Government.
  • Participants organize protests.
    • Posts are completely unlabeled (no pseudonyms)
    • Unknown organizational structure, number of authors, etc.
  • The government applies unsupervised stylometric techniques.
    • Number of authors may be discovered, author profiles created.
    • Results fed into supervised stylometry system to identify individuals.
Protecting Privacy: Attacking Stylometry

- **Problem**
  - Can stylometry be attacked? How so? How Easily?

- **Evaluation**
  - In depth study on attacking multiple methods of Stylometry.

- **Conclusion**
  - Stylometry is very vulnerable to attack by inexperienced human adversaries.
  - Attacks can be used to protect privacy.
We’re Under Attack!

• Obfuscation Attack
  • An author attempts to write a document in such a way that their personal writing style will not be recognized.

• Imitation Attack
  • An author attempts to write a document such that the writing style will be recognized as that of another specific author.

• Translation Attack
  • Machine translation is used to translate a document to one or more languages and then back to the original language.
Study Setup & Format

• 3 representative methods of stylometry.
• 15 Individual Authors. Participation had three parts:
  • Submit 5000 words of pre-existing writing from a formal source.
  • Write a new 500 word passage as an obfuscation attack.
    • Task: Describe your neighborhood.
  • Write a new 500 word passage as an imitation attack.
    • Task: Imitate Cormac McCarthy, describe your day.
• Authors had no formal training or knowledge in linguistics or stylometry.
Imitating Cormac McCarthy

“On the far side of the river valley the road passed through a stark black burn. Charred and limbless trunks of trees stretching away on every side. Ash moving over the road and the sagging hands of blind wire strung from the blackened lightpoles whining thinly in the wind.”
Imitation Attack Examples

• “Light sliced through the blinds, and construction began in the adjacent apartment. The harsh cacophony crashed through the wall.”

• “Hot water in the mug. Brush in the mug. The blade read ‘Wilkinson Sword’ on the layered wax paper packaging.”

• “He fills the coffee pot with water, after cleaning out the putrid remains of yesterday's brew. The beans are in the freezer, he remembers.”
Methodology

• Tested 4 methods of Authorship Recognition
  • Neural Network w/ 9 Features
  • Support Vector Machine w/ 9 Features
  • Synonym-Based Method
  • Writeprints
• Cross-validated across randomly selected data from a set of 12 unique authors.
Training Set Accuracy

Baseline Accuracy

Accuracy

No. of Authors

1  2  3  4  8  12  13

Baseline Accuracy

NN
SVM
Synonym
Writeprints
Random
Obfuscation Attack Detection

Obfuscation Accuracy

Accuracy

No. of Authors

NN
SVM
Synonym
Writeprints
Random
Imitation Attack Detection

[Graph showing imitation accuracy with different numbers of authors and various methods (NN, SVM, Synonym, Writeprints, Random).]
Imitation Attack Success

![Graph showing Imitation Success](chart.png)
Translation Attack

Google Translate (13 Authors)

Accuracy

No. of Authors

Baseline (EN)  EN-DE-EN  EN-JA-EN  EN-DE-JA-EN

Accuracy (%):
0.00%  10.00%  20.00%  30.00%  40.00%  50.00%  60.00%  70.00%  80.00%  90.00%  100.00%

Methods:
- NN
- SVM
- Synonym
- Random
The End of Stylometry?

• Not the end of stylometry in sensitive areas
  • New methods should test for adversarial threats.

• Stylometry is useful, but can also present a threat to privacy.
  • Attacking stylometry to preserve privacy has high potential.

• Potential for arms race:
  • Developing attack-resistant methods of stylometry vs. creating new attacks to preserve privacy.
Geeks in Washington
A Technologist at the FTC

• My job: Technologist, Federal Trade Commission’s Division of Privacy and Identity Protection (DPIP)
• DPIP operates under Section 5 of the FTC Act:
  • “Unfair methods of competition in or affecting commerce, and unfair or deceptive acts or practices in or affecting commerce, are hereby declared unlawful.”
• Investigations are non-public until closed or settled.
A Technologist’s Role in Policy

• I am a technical adviser to the legal staff of DPIP.
  • Interact with parties under investigation to foster a technical understanding of the issues.
  • Research potential violations and suggest investigations.
  • Educate staff on current trends and technologies that may be of interest.
  • Help construct legal documents that are technically sound.
The FTC Privacy Report

• Industry self-regulation efforts “have failed to provide adequate and meaningful protection.”
  • Regulation and legislation are necessary.

• Proposed Framework
  • “Privacy By Design”
  • Do Not Track Mechanism
    • Online Behavioral Advertising
    • Technical Approaches
  • Reasonable access to data.
Thanks.

• PSAL: psal.cs.drexel.edu
  • We are always looking for motivated students as applicants or research collaborators!

• Contact Me.
  • mb553@drexel.edu / mbrennan@ftc.gov
  • @brennan_mike
  • www.mbrennan.net
Twitter Addendum: Wednesday

- #everlastingfriends
- Eclipse soundtrack
- Zimbabwe
- Data Plans
- Almagro
- #ifhiphopwashighschool
- Joran
- Oil Spill
- Drake album
- #idontappreciate

Recall and Precision
Twitter Addendum: Thursday

#alliwannaknowis
#thankyouforbeingafriend
Janey
Joran
OrgulhoRestart
#nawimstraight
Bud Selig
Gano Mexico
Jim Joyce
Rue McLanahan

Recall
Precision
Twitter Addendum: Friday

- #areyoureallystill
- Coach Wooden
- Chace Crawford arrested
- Splice
- #oilspill
- Eunhyuk
- ChooseRestart
- Spelling Bee
- Rachel Corrie
- #makemebetter

![Bar chart showing recall and precision for various hashtags and topics.](chart.png)