Learning Topical Social Sensor on Twitter

Authors
Affiliation Address 1
Address 2

Abstract

Twitter represents a massively distributed social sensor of a rich underlying topic space that drives its content generation. Yet Twitter content is so diverse, decentralized, and dynamic in nature, that it is hard to automatically aggregate this topical content. To address this need, we provide a novel way of learning topical social sensors on Twitter that learn from a provided set of topical hashtags and generalize to identify topical tweets with previously unseen tags. These learning social sensors leverage a variety of user-based, hashtag-based, term-based, and location-based features for distinguishing topical from non-topical tweets; we further analyze these features to understand which features are most useful and why. We further assess general global topical trends and how our learning sensors are able to follow these trends by drawing from a rich variety of sources on the Twittersphere to enable a first generation of learning social sensors for Twitter.

Introduction

- Twitter is a vast sensor of content generated by latent phenonema (e.g., flu, political sentiment, elections, environment).
- Learning topical social sensors (politicians in NY, road conditions in Toronto) – very broad topics for which its hard to manually specify a useful query.
- But there is interesting topical content and wouldn't it be cool if we could learn a social sensor for a targeted topic?
- Key insight is that hashtags are topical and can be used to bootstrap a supervised learning system that as we will show generalizes well beyond the seed hashtags.
- Conclusion is a new way to build topical real-time feeds that are otherwise difficult to do with existing Twitter tools (???).

sectionLearning Topical Social Sensors

Start off with the questions that we want to answer in this section:

- How to evaluate, labeling (problem of no supervised labels for tweets, indirect via hashtags as topical surrogates, leads to question of hashtag curation)?

- Which classification algorithm is best / most robust for learning topical social sensors?

Dataset Statistics

- # tweets (by month – histogram)

- # users, #hashtags, obligatory power law plots of user tweet count, hashtag count

- different feature types and numbers (e.g., overall US and international location distribution choropleth) including total frequency counts of type across data (?)

- table of topics: 5 sample diverse training hashtags and #train/test hashtags, #tweets per topic

Experimental Methodology

How we curated hashtags: need to make up good story here. Inner-annotator agreement of 3/4.

Train/validation/test split date selection – temporally (.5,.1,.4)

Feature selection: threshold per feature 159 and 50 (just explain rationale for lower hashtag and location thresholds).

Formal notation, how do we train/test and tune hyperparameters for a generic classifier.

Classification Algorithms

1. Naive Bayes
2. Rocchio (centroid)
3. Logistic Regression

All above over 1,000,000 features, *same* training data for all algorithms.

Not breaking down by feature type yet – that's for the feature analysis section.

Analysis

- Table of rows:alg, cols: MAP, P@k (k in 10,100,1000) with stderrs over all topics

- Could do a bar graph (below) each for MAP, P@100 with topics as major columns and algs as neighboring bars

- Anecdotal results for each topic – point out deficiency in our labels (a good thing, we generalized well from small hashtag set), manual evaluation of relevance for top-100 for best algorithm?

naturaldisaster	**epidemics**	**irandeal**	**socialissues**	**lbgt**	**humancauseddisaster**	**celebritydeath**	**space**	**tennis**	**soccer**
philippines	usa	tehran	st.louis	usa	malaysia	southafrica	germany	london	liverpool
ca	ncusa	u.s.a	mo	bordentown	palestine	johannesburg	roodepoort	uk	manchester
india	garlandtx	nederland	usa	newjersey	syria	capetown	houston	india	london
newdelhi	oh-sandiego	iran	dc	sweethomealabama!	israel	pretoria	austin	pakistan	nigeria
newzealand	washington	globalcitizen	washington	aurora	london	durban	tx	islamabad	india
manila	dc	france	missouri	tennessee	pakistan	nairobi	virtualworld	mumbai	uk
wellington	smyrna	washington	brooklyn	co	kualalumpur	canada	sanfransisco	themidlands	anfield
sanfrancisco	newyork	londan	ny	nevada	gaza	kenya	ca	bangalore	newcastleupontyne
losangeles	chicago	london	saintlouis	unitedstatesofamerica	nigeria	gauteng	usa	england	lagos
uk	southernnewjersey	u.k	ca	sweethomealabama	washington	indonesia	oh-sandiego	melbourne	newcastle

Table 1: Top 10 Topics vs Locations based on Mutual Information

Feature Analysis

What we have to work with: topics, features, feature attributes

Questions/answers:

- **What are the best features for learning social sensors, do they differ by topic? (Why?)**
  1. Feature analysis aggregated over all topics:

     - Table: rows=features, cols=topics, table entries=Average Feature MI for that topic, final column for avg +/- stderr (might be better viewed in a colorized matrix)
  2. Feature analysis by topic:

     - Overall - 5 Boxplots: MI for 5 feature types vs. topics

     - Location analysis - Topic location *MI's*, topic *location frequencies* in boxplots (top-10 locations) or choropleths (need to avoid 10 choropleths, so need to have a way to pull out which topics might be interesting for location – can either select directly or use (i) to find which topics had high location MIs)
- **For each feature type, do any attributes correlate with importance?**

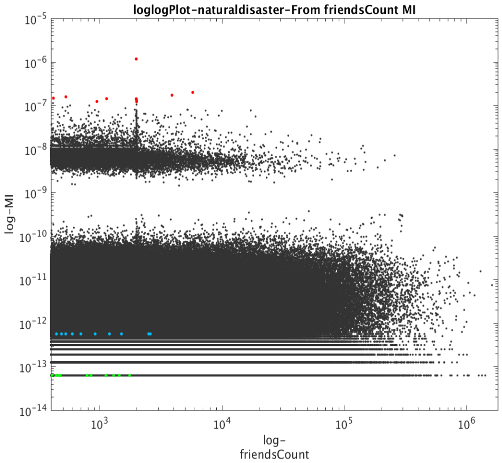

Figure 1: A sample scatter plot for Mutual Inforamtion of $FromUser$ vs. Friends Count (.eps format) that has been resized with the `epsfig` command.

  1. Anecdotal feature analysis: for each of 5 feature types: (rows) top-k / median-k (?), (cols) topics – much better than below b/c we show all topics here and we can compare features across topics.

     Don't use for now: show (rows) top-k and median-k features for different topics and (cols) 5 features (location, mention, from, term, hashtag) – need to select a **few (2-3) interesting topics** and explain shown in table 1
  2. scatterplots of feature MI – the absolute last thing we do (density plots?!!)

     **which plots below, and for which topics? Could pick out most useful features for topics in part (a)(i) and just show selected scatter plots below for these feature types.

     from, mention MIs vs. followers, favorites, friends, hashtags, tweets

     hashtag MI vs. #tweets, #users location MI vs. #users term MI vs. #tweets

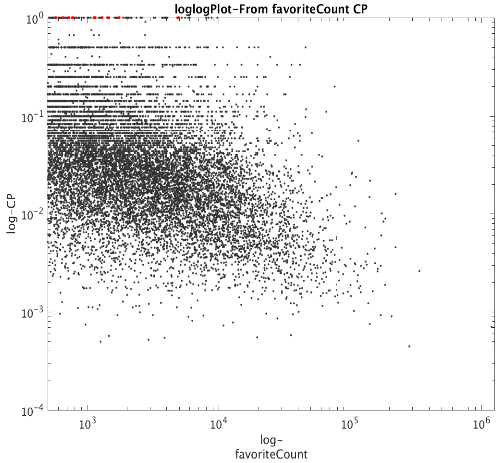

Figure 2: A sample scatter plot for Conditional Property of $FromUser$ vs. Favorite Count (.eps format) that has been resized with the `epsfig` command.

Related Works

Conclusions

Acknowledgments

(**?**)

Copyright

References

The aaai.sty file includes a set of definitions for use in formatting references with BibTeX. These definitions make the bibliography style fairly close to the one specified below. To use these definitions, you also need the BibTeX style file “aaai.bst,” available in the author kit on the AAAI web site. Then, at the end of your paper but before \enddocument, you need to put the following lines:

```
\bibliographystyle{aaai} \bibliography{bibfile1,bibfile2,...}
```