LinkedIn™

Topic-Based Sentiment Analysis: Mining Member Feedback and Social Media for Actionable Insights

Vita Markman, Staff Software Engineer
& Yongzeng Zhang, Manager
Business Analytics

Introduction

What is this talk about?

Mining **raw text** for **key topic phrases**,
obtaining **sentiment** about the topic phrases,
and providing **word context** to explain it.

Topic based sentiment = sentiment attributed to a specific topic in a document vs. sentiment of a document as a whole

The Rationale

Reasons for topic-based sentiment analysis:

(1) **Sentiment-related:** actionable information requires nuance: what **specifically** do people like vs. dislike?

(2) **Content-related:** discovering a semantic landscape that characterizes a text corpus

Example: a collection of social media posts about "jobs" has a different topic distribution than one about "inbox".

The Challenges

Focus of this talk is on the following challenges:

1. Discovering **informative topics** in noisy text
2. Finding **sentiment** of the topics
3. **Explaining** topic **sentiment** via **word attributes** that offer more **context**

Data: publically available social networks posts mentioning LinkedIn grouped into macro categories via keywords like "Jobs" or "Inbox"

A Glimpse at Raw Data

The Next Big Thing You Missed**: LinkedIn**'s Quest to Get a **Job** for Everyone on Earth...

Research shows that 75% of **jobs** are found through #networking, compared with 10% through advertisements, 10% through recruiters and 5% through a job fair …

Understanding Networking and using #LinkedIn as a Tool to Help You Find a **Job**...

You'll receive resume and **LinkedIn** profile tips, **job** interview tips, **job** market insider stats, and the latest in career and business development resources. …

Topic: a Working Definition

A **topic** is an ngram size >=2 that respects the following conditions:

1. **syntactically** well formed phrase, an NP or VP
2. **semantically informative** w.r.t. a given **corpus**

Topic:		**Not a topic:**
talent solutions	vs.	solutions of
inmail policy	vs.	good day
invitation to connect	vs.	connect with

Two Methods of Topic Discovery

Method 1: define patterns of POS tag that carve out well formed phrases such as NP, VP, AP

Problem1: may not work well with noisy text

Problem2: may pick up semantically uninformative phrases like “very good” and “greatly appreciated”

Method 2 of obtaining topics:

- Make no reference to POS tags - suitable for noisy text.
- Eliminate semantically uninformative phrases (E.g. good day) .
- Eliminate syntactic fragments (E.g. CEO of)

Topic Extraction: an Illustration

Raw Ngrams from corpus:

The book,
On linkedIn,
In jobs,
Jobs on linkedin,
Recommendations and,
Good morning,
Is looking
Take care,
Career opportunities,
Business development,
CEO Jeff Weiner,
Hiring manager,
Manager of
Of your department
Opportunities in
In business
Talent solutions
….

Top frequent ~1000 words from any other corpus

"the"
"a"
"of"
"is"
..
"take"

Phrasal ngrams :
Career opportunities
Business development
Talent Solutions
Jobs on LinkedIn

Algorithm for Topic Extraction

1. Define externalWordList = top 1000 frequent words from an external corpus
2. Separate each social media post ngrams size >=2
3. *Foreach ngram in ngram_set:*

    words = tokenize(ngram)

    *if words[0] not in externalWordList *

    & words[-1] not in externalWordList:

        phrasalTopics.add(ngram)

phrasalTopics: {business development, hiring manager, profile on LinkedIn}

NOT phrasal & hence discarded: {development and, to hiring, good morning}

Obtaining Topic Sentiment

1. For each topic find all clauses containing it
2. Rate each clause for sentiment via SVM

Examples:

*LinkedIn is the worlds top **[professional network]**^{TOPIC}* *Positive*

*i use linkedin a lot for **[business development]**^{TOPIC}* *Positive*

3. Aggregate rated clauses to get "majority vote" for each topic

Examples:

Business development: positive 0.33, neutral 0.66

Recruiters use linkedin: positive: 0.66, neutral: 0.33

Social networking: positive: 0.60, neutral: 0.37, negative: 0.3

Recall: **~ 80%** [we are able to assign sentiment to ~80% of data]

Precision: ~ 70% [of the ones we label, we label correctly 70%

Attributes: Context for Topic Sentiment

Attributes = words that **appear in neighboring context** of the **topic** and help **explain** the **sentiment** of the topic.

We want to know **why** the sentiment for *recruiters use LinkedIn* **positive**

For each topic t_i find all the words that appear in the same clause as t_i

- Use tf*idf of each attribute to retain attributes informative of a given topic, e.g. rare across all topics
- Topic attributes can be used to find semantically similar topics and combine them together

Examples of Topic Attributes

recruiters use linkedin ['modern', 'hiring', 'identify’, 'candidate', 'source’, 'potential', 'interview', 'improve', 'executive', 'hires', 'resume', 'qualified’, 'employees', 'recruit’]

linkedin lead generation : ['engaging', 'prospects', 'create’, 'powerful', 'content', 'secret', 'branding', 'potential’]

networking tool : ['powerful’, 'elearning', 'startups', 'network', 'valuable', 'community’, 'friends', 'vibes', 'social', 'directory', 'inbox’]

business development ['executive', 'hiring', 'manager', 'planning', 'solutions', 'portfolio’, 'impact', 'openings’, 'division', 'updated’, 'execution']

A Semantic Landscape of a Corpus

Topics for "Jobs" Category

hiring manager
interview questions
linkedin recommendations
business development
professional network
talent solutions
talent acquisition
relationship manager
career opportunities

Topics for "Inbox" Category

networking tool
linkedin lead generation
linkedin invitation template
social network
sending messages
inmail credits
inmail on linkedin
sending multiple emails
friend request

Different Collections, Different Landscapes

Topics in “jobs” mention *resume, career, hiring, and talent*

Topics in “inbox” mention *invitations*, *inMail, and leads*

Frequent topics in “jobs” collection e.g. *talent solutions* or *career opportunities* are entirely absent in “inbox” collection.

Some ‘general’ topics overlap between the two – *social network* and *professional network*

Discovered top topics for each category can “summarize” the overall content of the corpus

Summary

The proposed topic-based sentiment mining approach :

- Works well on the language of social media
- Removes uninformative or fragmented phrases e.g. “good morning” or “CEO of” via leveraging an external corpus
- Leads to actionable insights via the more nuanced topic-based sentiment
- Discovers attributes to explain the sentiment of a topic.
- Offers **a semantic landscape** of a collection of text via discovered phrasal topics.
- Can be leveraged to **summarize** a variety of **corpora** regardless of the noisiness of the raw data

Thank You!