Lecture 12: Sentiment Analysis

Pilsung Kang
School of Industrial Management Engineering
Korea University

AGENDA

Definition

Feldman (2013)

- What is Sentiment Analysis?
  - ✓ Computational study of opinions, sentiments, evaluations, attitudes, appraisal, affects, views, emotions, subjectivity, etc., expressed in text.
    - Text: Reviews, blogs, discussions, news, comments, feedback,…
  - ✓ Sentiment Analysis is sometimes called **Opinion Mining**

Backgrounds

Feldman (2013)

- Typical Usage of Sentiment Analysis
  - ✓ Extract from text how people feel about different products
  - ✓ Sentiment analysis can be tricky
    - **`Honda Accords and Toyota Camrys are nice sedans`**
    - **`Honda Accords and Toyota Camrys are nice sedans, but hardly the best car on the road`**

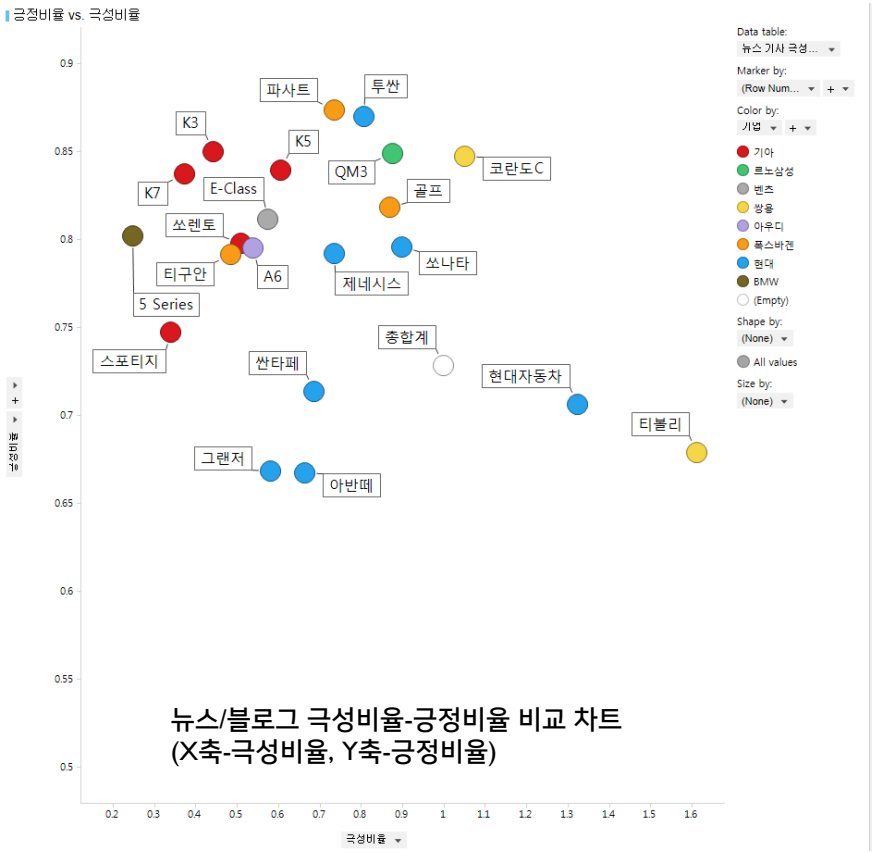
Backgrounds

Feldman (2013)

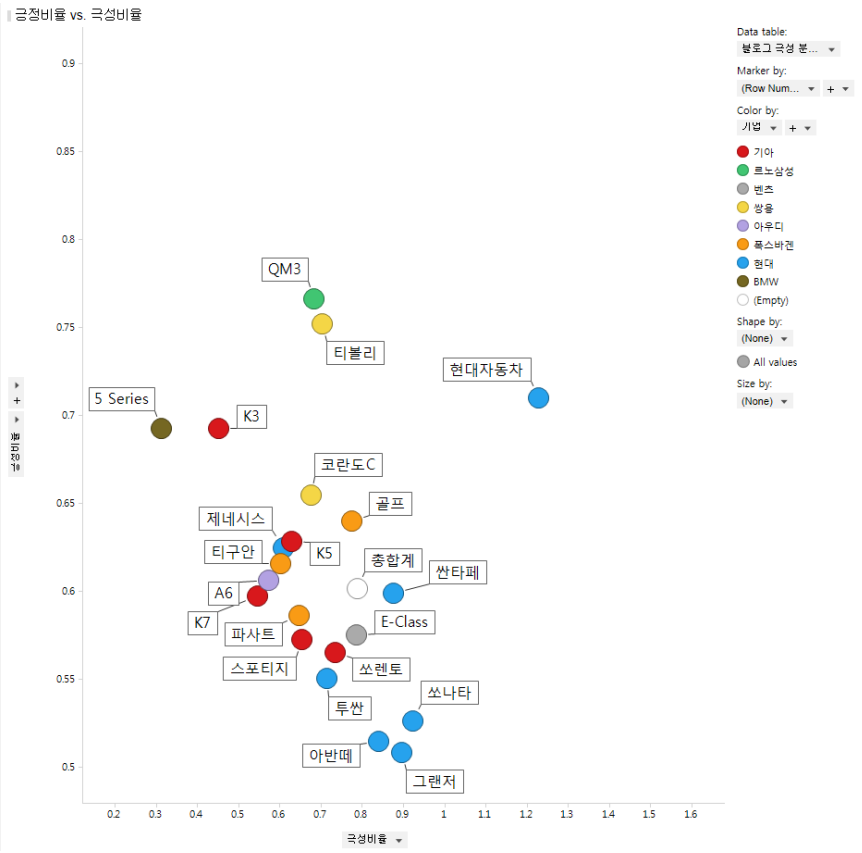
- Opinions are widely stated
  - ✓ Organization's internal data
    - Customer feedback from emails, call centers, etc.
  - ✓ News and reports
    - Opinions in news articles and commentaries
  - ✓ Word-of-mouth on the web
    - Personal experiences and opinions about anything in reviews, forums, blogs, Twitter, micro-blogs, etc.
    - Comments about articles, issues, topics, reviews, etc.
    - Posting at social networking sites (Facebook, Instagram, etc.)

Applications

- Business and organizations
  - ✓ Benchmark products and services; market intelligence
    - Business spend a huge amount of money to find customer opinions using consultants, surveys and focus groups, etc.



뉴스/블로그 극성비율-긍정비율 비교 차트
(X축-극성비율, Y축-긍정비율)



Applications

- Individuals
  - ✓Make decisions to purchase products or to use services
  - ✓Find public opinions about political candidates and issues

Applications

- Ad placement
  - ✓ Place an ad if one praises a product
  - ✓ Place an ad from a competitor if one criticizes a product

Guest

Why sentiment analysis is the future of ad optimization

PETER YARED MARCH 20, 2011 6:21 PM

TAGS: KLOUT, PEERINDEX

[Peter Yared is the vice presi
acquired Transpond, a socia

Sentiment analysis is a hot
the promise of helping bra
are thinking and saying ab
including early contender
BuzzLogic, and my own co
Measurement product are becoming pervasive
while consumer sentiment is important, what's
revenue.

First Workshop
on

Internet Advertising Using Sentiment Analysis (AdSent 2013)
with

IEEE International Conference on Data Mining series (ICDM), December 7, Dallas, Texas, USA

Applications

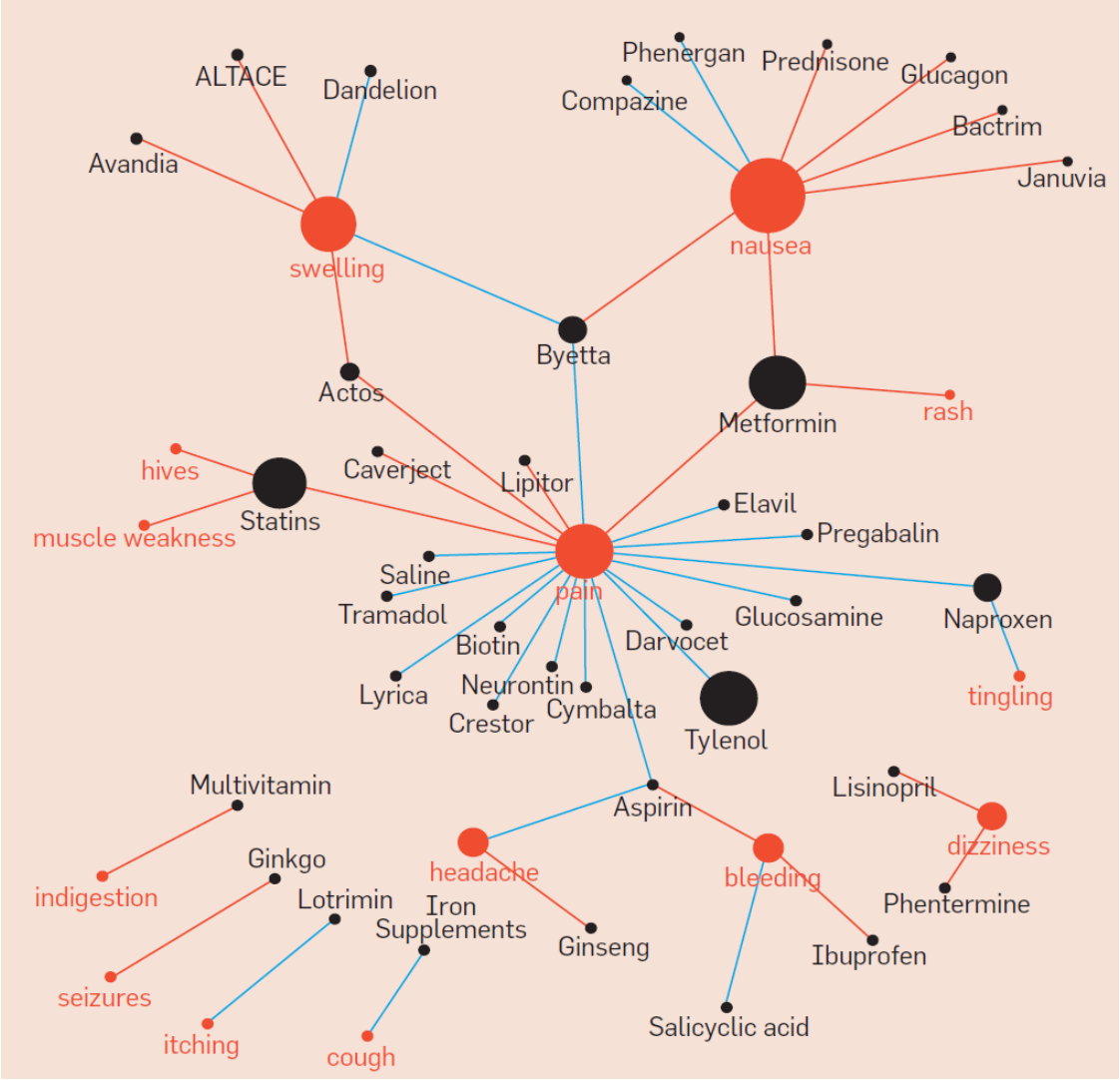
- Opinion Retrieval
  - ✓ Provide general search for opinions

Player	Line	Score	승	패배	피안타	피홈런	볼넷	실점	평균 자책점	이닝당 삼진	이닝당 볼넷	삼진 /볼넷	방어율
웨버	NC웨버는 7이닝 동안 1안타 3볼넷 10탈삼진으로 무실점 호투했다.	0.991	1	0	0	0	1	0	0	1	1	1	0
이재학	이재학 '10승', 3년만에 달라진 위상 증명하다	0.986	1	0	0	1	0	1	0	0	0	1	0
이태양	'이태양 완벽투' NC, LG 상대로 창단 첫 스윕	0.966	1	0	0	0	0	0	0	0	0	1	0
에릭	한편, 이날 7이닝 3실점으로 잘 던지며 7경기 만에 국내 첫 승을 거둔 에릭은 "굉장히 흥분된다.	0.803	1	0	1	0	0	0	0	0	0	0	0
찰리	어이없는 실책 2개가 나오자 찰리는 흔들리기 시작했다.	0.506	0	1	1	1	0	1	0	1	0	1	1
찰리	찰리가 대량 실점을 했지만 그 과정에서 3루수 모창민과 1루수 테임즈의 실책이 동반되면서 자책은 1점 밖에 되지 않았다.	0.506	0	1	1	1	0	1	0	1	0	1	1
원종현	NC 원종현이 패전투수가 됐다.	0.293	0	1	0	0	0	1	1	1	0	1	1
원종현	세 번째 투수로 등판한 원종현 역시 1사 후 8번 김성현에게 솔로포를 맞았다.	0.102	0	0	0	1	1	0	1	1	1	0	0
이재학	NC 선발 이재학은 8이닝 8피안타(2홈런) 5탈삼진 3볼넷 2실점을 기록했다.	0.079	0	0	1	1	0	1	1	0	0	0	1
이혜천	이혜천이 올라온 뒤 한꺼번에 5점을 내주면서 맥빠진 경기가 되고 말았다.	0.066	0	0	1	1	0	1	0	0	1	0	0

Applications

Feldman (2013)

- Relationships between Drugs and Symptoms



AGENDA

Problem Definition

Feldman (2013)

Problem Statement: Abstraction

- Consists of two parts
  - ✓ (1) Opinion definition
    - What is an opinion?
  - ✓ (2) Opinion summarization
    - Opinions are subjective
    - An opinion from a single person (unless a VIP) is often not sufficient for action
    - We need opinions from many people, and thus need to summarize those opinions

Problem Definition

Feldman (2013)

Entity and Aspect/Feature

> **Id: Abc123 on 5-1-2008** "*I bought an iPhone a few days ago. It is such a nice phone. The touch screen is really cool. The voice quality is clear too. It is much better than my old Blackberry, which was a terrible phone and so difficult to type with its tiny keys. However, my mother was mad with me as I did not tell her before I bought the phone. She also thought the phone was too expensive, …*"

- What do we see?
  - ✓ Opinion targets: entities and their features/aspects
  - ✓ Sentiments: positive and negative
  - ✓ Opinion holders: persons who hold the opinions
  - ✓ Time: when opinions are expressed

Problem Definition

Feldman (2013)

- Two main types of opinions
  - ✓ Regular opinions: Sentiment/opinion expressions on some target entities
    - Direct opinions: "The touch screen is really cool."
    - Indirect opinions: "After taking the drug, my pain has gone."
  - ✓ Comparative opinions: Comparisons of more than one entity
    - "iPhone X is better than Galaxy 9"
  - ✓ Focus more on regular opinions in this lecture

Problem Definition

Feldman (2013)

- A (Regular) Opinion
  - ✓ An opinion has the following basic components

$$(g_i, so_{ijk}, h_j, t_k)$$

where

- g_i is a target.
- so_{ijk} is the sentiment value (positive, negative, neutral, or a rating score) of the opinion from opinion holder h_j on target g_i at time t_k.
- h_j is an opinion holder.
- t_k is the time when the opinion is expressed.

Problem Definition

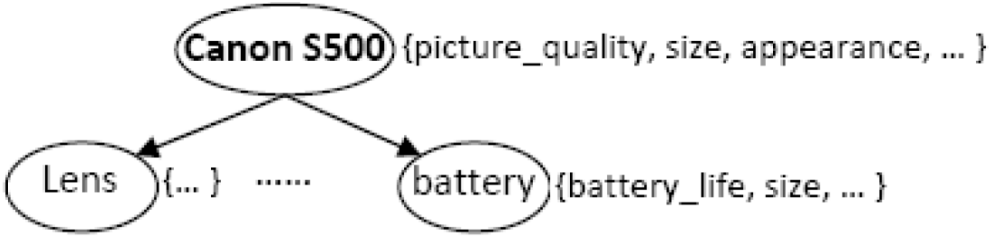
Feldman (2013)

- Opinion Target
  - ✓ In some cases, opinion target is a single entity or topic.
    - "I love G5" and "I support tax cut."
  - ✓ But in many other cases, it is more complex.
    - "I bought an iPhone a few days ago. It is such a nice phone. The touch screen is really cool."
    - Opinion target of the 3rd sentence is not just touch screen, but the "touch screen of iPhone"
    - "I support tax cut for the middle class, but not for the rich…"

Problem Definition

Feldman (2013)

- Entity and Aspect
  - ✓ Definition of Entity: An **entity e** is a product, person, event, organization, or topic. **e** is represented as
    - a hierarchy of **components**, **sub-components**, and so on.
    - Each node represents a component and is associated with a set of **attributes** of the component.



    - An opinion can be expressed on any node or attribute of the node.
    - For simplicity, we use the term **aspects (features)** to represent both components and attributes.

Problem Definition

Feldman (2013)

- Definition of an Opinion as a Quintuple

$$(e_i, a_{jl}, so_{ijkl}, h_j, t_k)$$

where

- e_i is a target entity/object
- a_{jl} is an aspect/feature/attribute/facet of the entity e_i
- so_{ijkl} is the sentiment value (positive, negative, neutral, or a rating score) of the opinion from opinion holder (source) h_j on aspect a_{jk} of entity e_i at time t_k
- h_j is an opinion holder
- t_k is the time when the opinion is expressed

Problem Definition

Feldman (2013)

- Quintuple Opinion Examples

> **Id: Abc123 on 5-1-2008** *"I bought an iPhone a few days ago. It is such a nice phone. The touch screen is really cool. The voice quality is clear too. It is much better than my old Blackberry, which was a terrible phone and so difficult to type with its tiny keys. However, my mother was mad with me as I did not tell her before I bought the phone. She also thought the phone was too expensive, …"*

✓ (iPhone, General, Positive, Abc123, 5-1-2008)

✓ (iPhone, touch_screen, Positive, Abc123, 5-1-2008)

✓ …

Problem Definition

Feldman (2013)

Purpose: Structure the unstructured

- Given an opinionated document,
  - ✓ Discover all quintuples $(e_i, a_{jl}, so_{ijkl}, h_j, t_k)$
  - ✓ Or, solve some simpler form of the problem
    - E.g. Classify the sentiment of the entire document
  - ✓ With the quintuples, unstructured texts transformed into structured data.

Rational vs. Emotional Evaluation

Feldman (2013)

- Rational Evaluation
  - ✓ Many evaluation/opinion sentences express no emotion
    - The voice of this phone is clear
- Emotional evaluation
    - I love this phone
- Some emotion sentences express no (positive or negative) opinion/sentiment
  - ✓ I am so surprised to see you

Abstraction: Opinion Summary

Feldman (2013)

- Feature-based opinion summary

""I bought an iPhone a few days ago. It is such a nice phone. The touch screen is really cool. The voice quality is clear too. It is much better than my old Blackberry, which was a terrible phone and so difficult to type with its tiny keys. However, my mother was mad with me as I did not tell her before I bought the phone. She also thought the phone was too expensive, …"

1.

….

Feature Based Summary of iPhone:

Feature1: **Touch screen**

Positive: 212

- *The touch screen was really cool.*
- *The touch screen was so easy to use and can do amazing things.*

…

Negative: 6

- The screen is easily scratched.
- I have a lot of difficulty in removing finger marks from the touch screen.

…

Feature2: **voice quality**

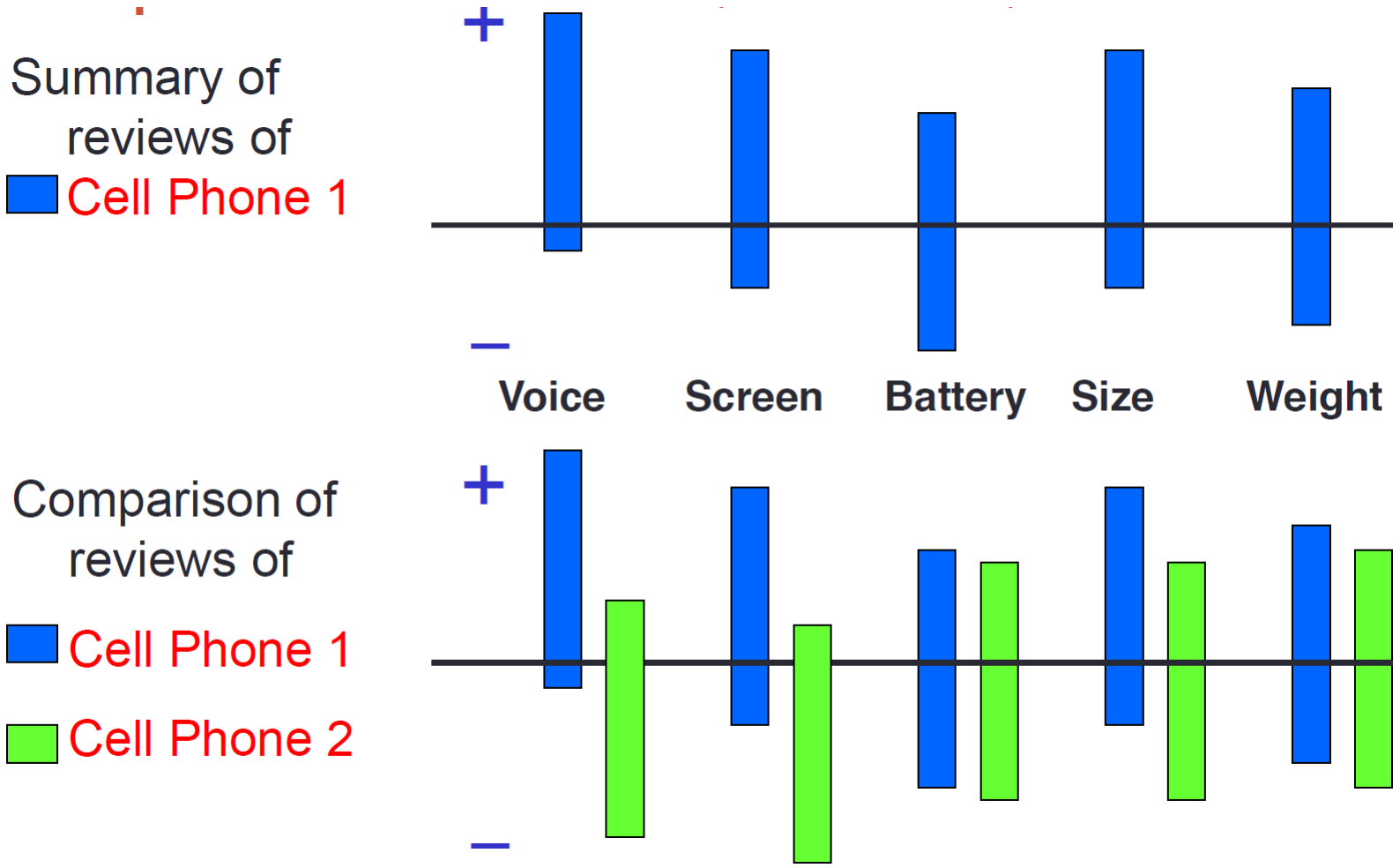
…

Note: We omit opinion holders

Abstraction: Opinion Summary

Feldman (2013)

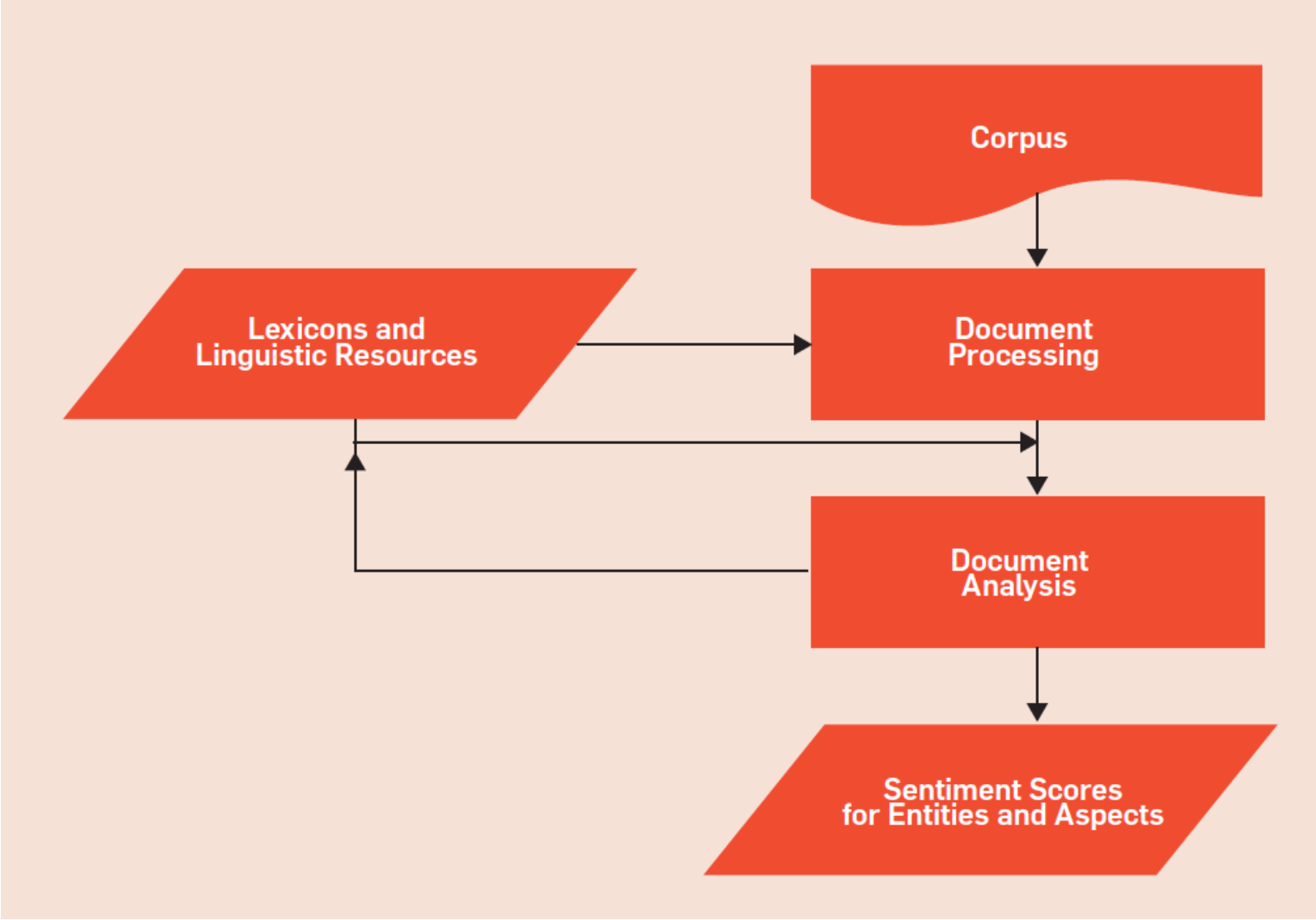
- Opinion Comparison



Sentiment Analysis Architecture

Feldman (2013)

- Architecture of generic sentiment analysis system



Document Sentiment Classification

Feldman (2013)

- Document Sentiment Classification
  - ✓ Classify a whole opinion document based on the overall sentiment of the opinion holder
    - Classes: **positive**, **negative** (possibly neutral)
    - Neutral or no opinion is hard. Most papers ignore it.
  - ✓ It is basically a text classification problem
  - ✓ Assumption: a document is written by **a single person** and expresses opinion/sentiment on **a single entity**.
  - ✓ Goal: discover **(_ , _ , so, _, _)** where e, a, h, and t are ignored.
    - Reviews usually satisfy the assumption (4/5 stars → positive, ½ stars → negative)
    - Many forum postings and blogs do not. They can mention and compare multiple entities or they express no sentiments.

Document Sentiment Classification

Feldman (2013)

- Supervised Learning Formulation
  - ✓ Training and test data: movie reviews with star ratings
    - 4-5 stars as positive and 1-2 stars as negative
    - Neutral is ignored
  - ✓ **Key: feature engineering**
    - Typically unigrams (bag of individual words)
    - Term frequency and different IR weighting schemes
    - Part of speech (POS) tags
    - Opinion words and phrases
    - Negations
    - Syntactic dependency
  - ✓ Too coarse for most applications

Sentence Sentiment Classification

Feldman (2013)

- Sentence Sentiment Classification
  - ✓ Assumes a single sentiment per sentence
  - ✓ Not always true, so one can classify clauses instead

- Two steps of sentence sentiment classification
  - ✓ Subjectivity classification
    - To identify subjective sentences
  - ✓ Sentence classification of subjective sentences
    - As positive or negative

Feature/Aspect-based Sentiment Classification

Feldman (2013)

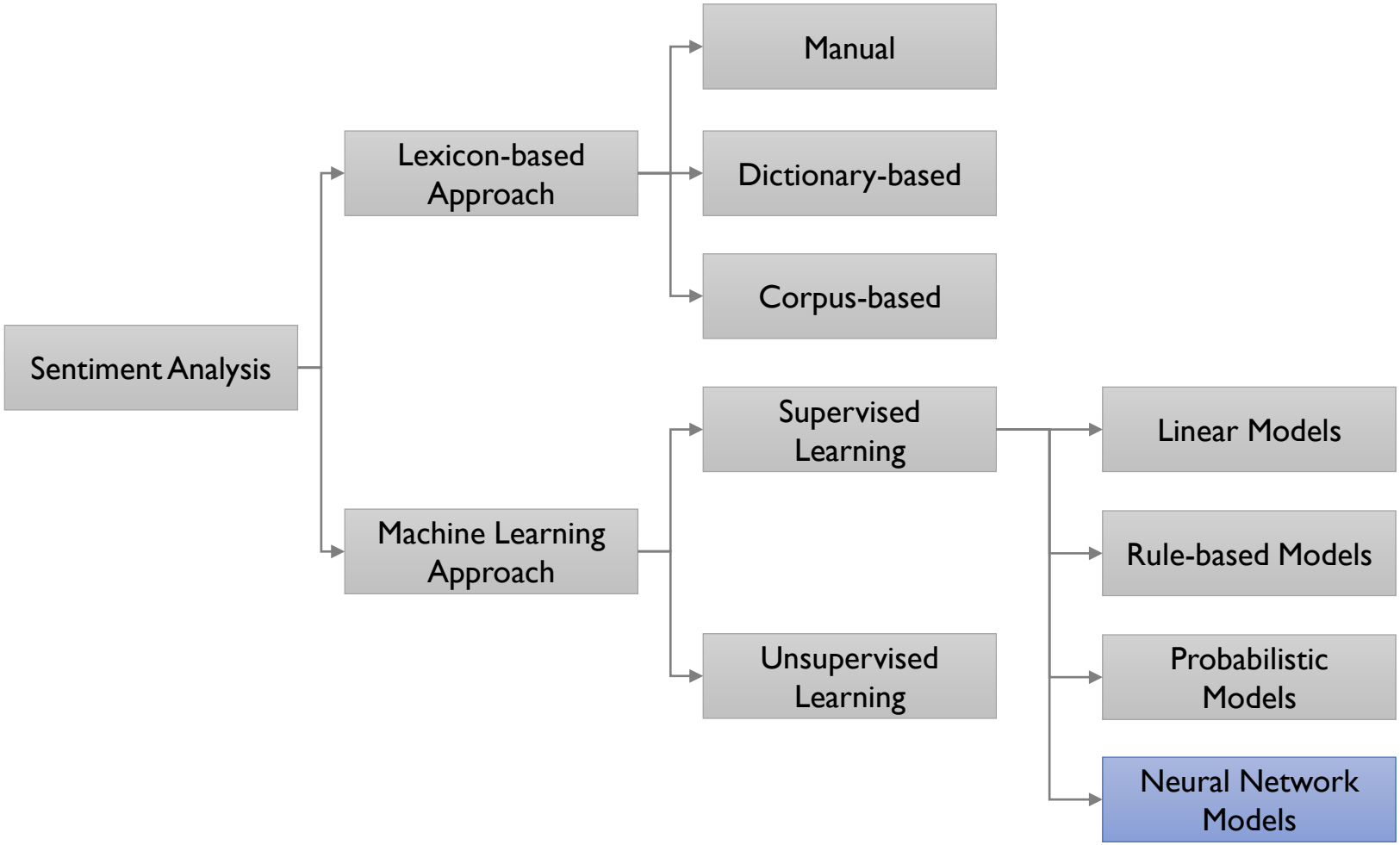
- Document/Sentence level sentiment classification
  - ✓ Useful but do not find what people liked and disliked.
  - ✓ They do not identify the targets of opinions: entities and their aspects
- Finding entities
  - ✓ Similar to but somewhat different from the traditional named entity recognition (NER)
    - Frequent nouns and noun phrases
    - Rule-based or supervised learning (HMM, CRFs)
    - Double propagation (DP): knowing one helps find the other (an opinion should have a target)
      - Input: a set of seed opinion words
      - The rooms are spacious
      - The phone has good screen

Sentiment Lexicon

- Sentiment Lexicon
  - ✓ Lists of words and expressions used to express people's subjective feelings and sentiment/opinions
    - Positive: beautiful, wonderful, good, amazing, …
    - Negative: bad, poor, terrible, …
    - Not just individual words, but also phrases and idioms
    - "cost an arm and a leg"
  - ✓ Many of them are context dependent, not just application domain dependent.
- Three main ways to compile such lists
  - ✓ Manual approach: not a bad idea for a one-time effort
  - ✓ Dictionary-based approach
  - ✓ Corpus-based approach

Sentiment Analysis Approaches

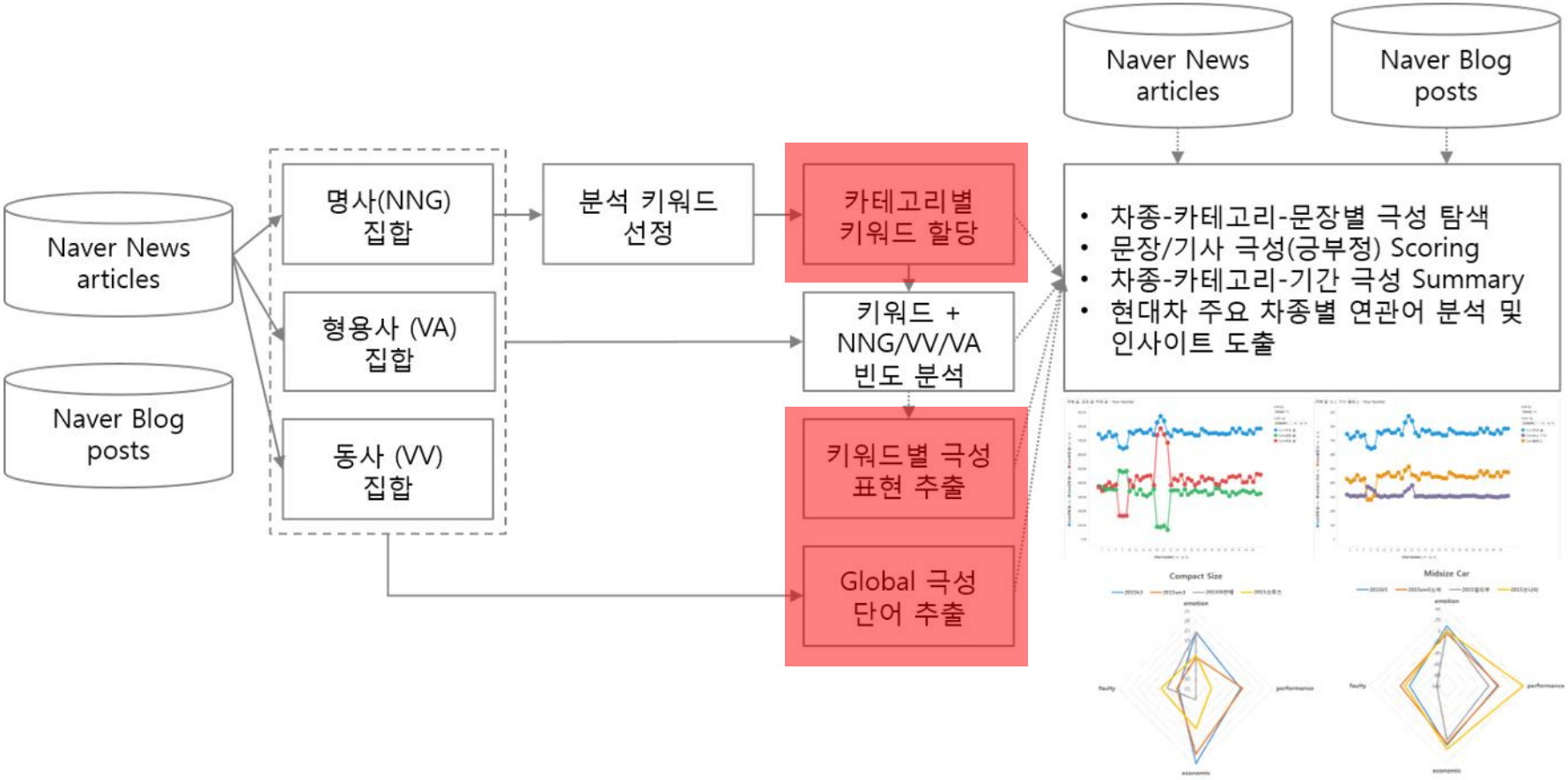
- Sentiment classification techniques



AGENDA

Manual Approach

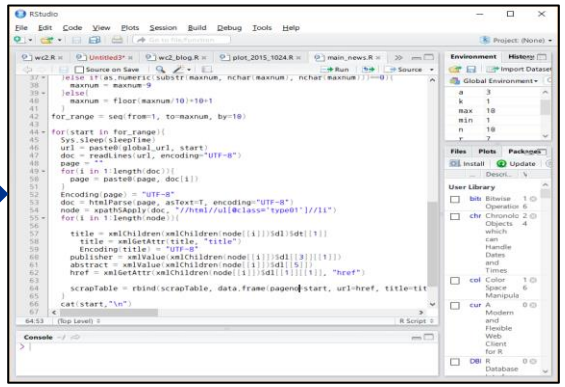
- Sentiment Analysis for Car Models



Manual Approach

- Data Collection
  - ✓ News articles and blog posts from NAVER (2010. 01. 01 ~ 2015. 07. 31)

원본 기사/블로그 게시글 검색



웹 크롤러를 이용한 게시글 수집

article_date	title	sentence_id	sentence
2010-01-04	2009 `베스트셀링카'는 쏘나타	2	(서울=연합뉴스) 권혁창 기자 = 지난해 국내에서 가장 많이 팔린 베스트셀링카는 현대차의 쏘나타로 나타났다
2010-01-04	2009 `베스트셀링카'는 쏘나타	3	4일 관련업계에 따르면 현대차 쏘나타는 지난 한해 구형 NF 8만4천981대, 신형 YF 6만1천345대 등 총 14만6천326대가 팔려 준중형 아반떼(11만5천378대)를 누르고 내수 판매 1위를 차지했다
2010-01-04	2009 `베스트셀링카'는 쏘나타	4	특히 신형 쏘나타는 지난해 9월 출시된 이후 매달 1만5천대 이상 팔려나가며 현대차의 내수판매 증대를 이끌었다
2010-01-04	2009 `베스트셀링카'는 쏘나타	5	기아차의 경차 모닝은 10만2천82대로 3위에 올랐으며, 현대차의 소형트럭 포터가 7만8천846대로 4위에 랭크됐다
2010-01-04	2009 `베스트셀링카'는 쏘나타	6	이어 현대차 그랜저(7만5천844대), 싼타페(5만8천324대), 르노삼성 SM5(6만1천10대)가 5~7위를 차지했다
2010-01-04	2009 `베스트셀링카'는 쏘나타	7	기아차 포르테(5만1천374대), 로체(4만9천54대), 르노삼성 SM3(4만5천906대), GM대우 라세티 프리미어(4만4천464대)가 그 뒤를 이었다
2010-01-04	2009 `베스트셀링카'는 쏘나타	8	faith@yna.co.kr
2010-01-04	2009 `베스트셀링카'는 쏘나타	9	[관련기사]
2010-01-04	2009 `베스트셀링카'는 쏘나타	10	<저작권자(c)연합뉴스
2010-01-04	2009 `베스트셀링카'는 쏘나타	11	무단전재-재배포금지.>

데이터베이스에 Raw Data를 저장

Manual Approach

- Data Collection: 520,000 news articles & 2,360,000 blog posts
  - ✓ At least 10,000 and at most 38,000 news articles are collected for each model
  - ✓ At least 89,000 and at most 248,000 blog posts are collected for each model

원본 문장	형태소 분석 결과
지난해 국내에서 가장 많이 팔린 베스트셀링카는 현대차의 쏘나타로 나타났다.	지난해(NNG) 국내(NNG) 팔리(VV) 베스트(NNG) 셀링카(NNG) 현대(NNG) 차(NNG) 쏘나타(NNG) 나타나(VV)
4일 관련업계에 따르면 현대차 쏘나타는 지난 한해 구형 NF 8만4천981대, 신형 YF 6만1천345대 등 총 14만6천326대가 팔려 준중형 아반떼(11만5천378대)를 누르고 내수 판매 1위를 차지했다.	관련(NNG) 업계(NNG) 따르(VV) 현대(NNG) 차(NNG) 쏘나타(NNG) 지나(VV) 한해(NNG) 구형(NNG) 신형(NNG) 팔리(VV) 중형(NNG) 아반떼(NNG) 누르(VV) 내수(NNG) 판매(NNG) 차지(NNG)
특히 신형 쏘나타는 지난해 9월 출시된 이후 매달 1만5천대 이상 팔려나가며 현대차의 내수판매 증대를 이끌었다.	신(NNG) 형(NNG) 쏘나타(NNG) 지난해(NNG) 출시(NNG) 이후(NNG) 매달(NNG) 이상(NNG) 팔(VV) 나가(VV) 현대(NNG) 차(NNG) 내수(NNG) 판매(NNG) 증대(NNG) 이끌(VV)
기아차의 경차 모닝은 10만2천82대로 3위에 올랐으며, 현대차의 소형 트럭 포터가 7만8천846대로 4위에 랭크됐다.	기아(NNG) 차(NNG) 경차(NNG) 모닝(NNG) 오르(VV) 현대(NNG) 차(NNG) 소형(NNG) 트럭(NNG) 포터(NNG) 랭크(NNG)
이어 현대차 그랜저(7만5천844대), 싼타페(5만8천324대), 르노삼성 SM5(6만1천10대)가 5~7위를 차지했다.	잇(VV) 현대(NNG) 차(NNG) 그랜저(NNG) 싸(VA) 타(NNG) 르노(NNG) 삼성(NNG) SM5(NNG) 가(VV) 차지(NNG)
기아차 포르테(5만1천374대), 로체(4만9천54대), 르노삼성 SM3(4만5천906대), GM대우 라세티 프리미어(4만4천464대)가 그 뒤를 이었다.	기아(NNG) 차(NNG) 포르테(NNG) 로체(NNG) 르노(NNG) 삼성(NNG) SM3(NNG) 대우(NNG) 라세티(NNG) 프리미어(NNG) 가(VV) 뒤(NNG) 잇(VV)

Manual Approach

- Keyword selection and sentiment dictionary construction

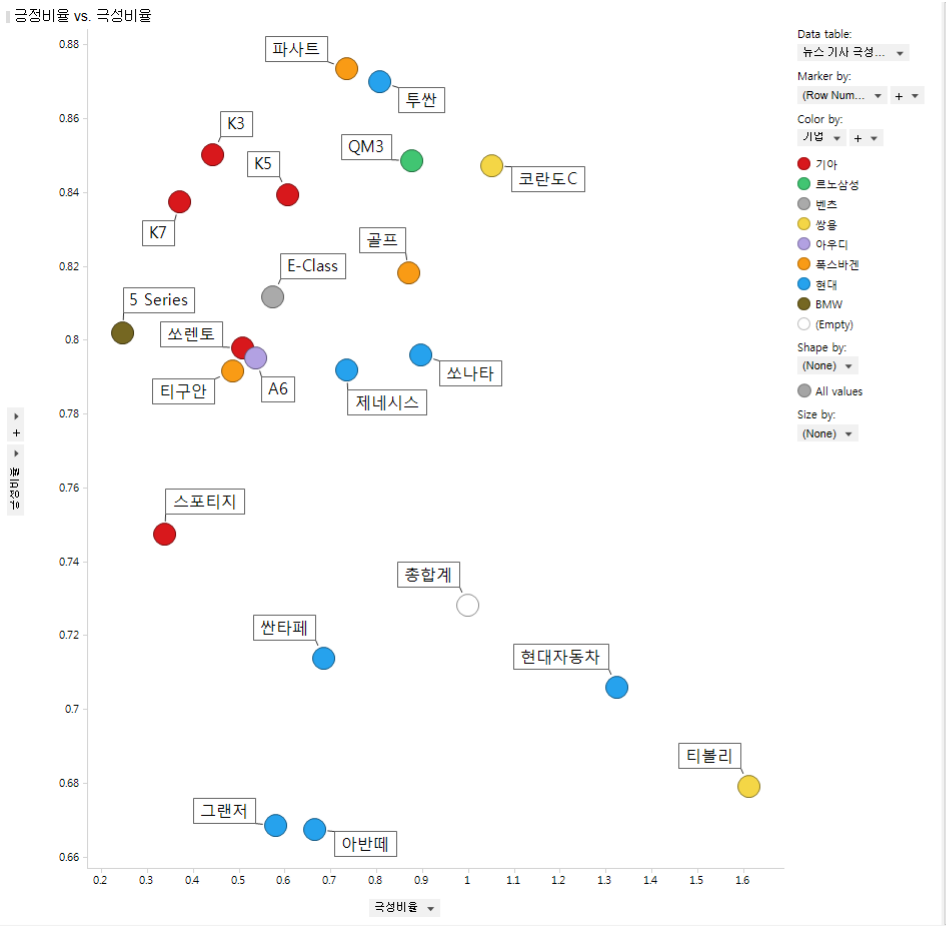
  ✓ Purely done by X people within 7 days

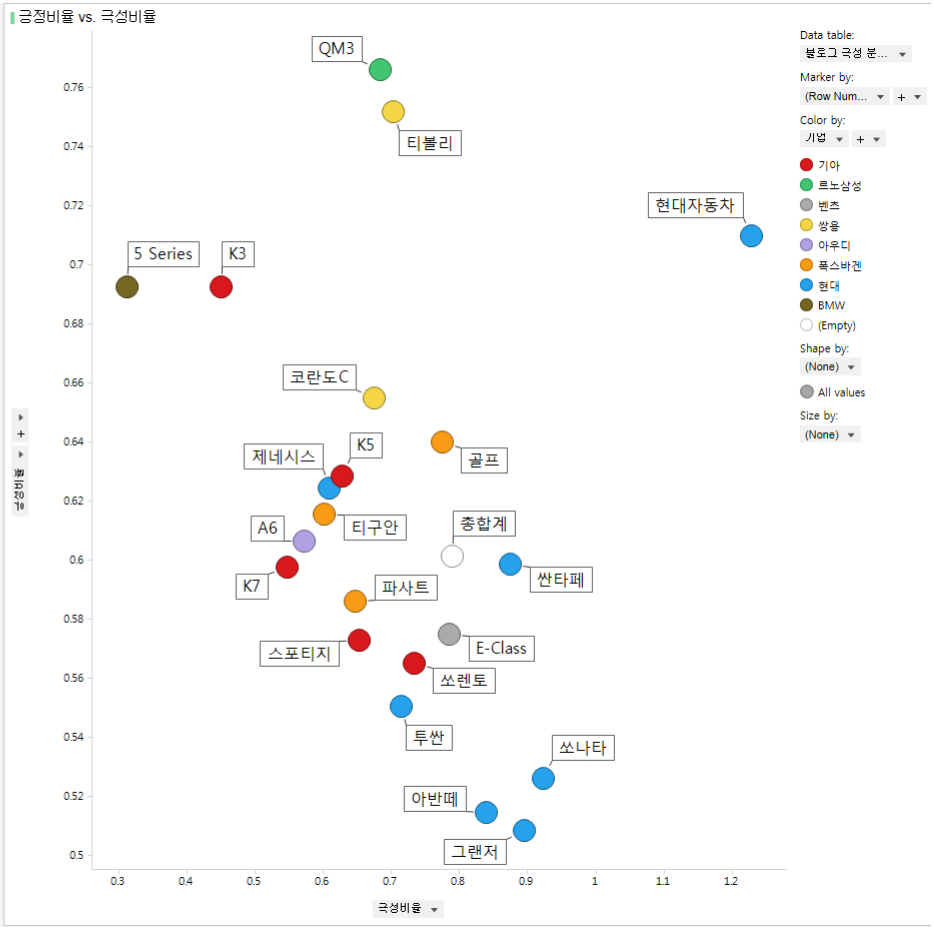
카테고리	키워드	키워드+표현	극성
Performance	기술	기술 갖(VV)	Positive
Economic	시장	시장 티볼리(NNG)	Negative
Reputation	노사	노사 연속 파업(NNG)	Negative
Performance	엔진	엔진 선택 있(VV)	Positive
Economic	매출	매출 하락(NNG)	Negative
Reputation	직원	직원 노동 조합(NNG)	Negative
Performance	하이브리드	하이브리드 고객(NNG)	Positive
Faulty	경보	경보 장치 급제동(NNG)	Positive
Performance	마력	마력 엔진 장착(NNG)	Positive
Economic	경영	경영 환경 악화(NNG)	Negative
Emotional	프리미엄	프리미엄 이미지(NNG)	Positive
Emotional	사양	사양 정숙(NNG)	Positive
Economic	시장	시장 다변화(NNG)	Positive
Emotional	출시	출시 스포츠(NNG)	Positive
Economic	시장	시장 판매 증가율(NNG)	Positive
Economic	환경	환경 차량 선보이(VV)	Positive
Emotional	콘셉트	콘셉트 세계 최초(NNG)	Positive
Reputation	노동자	노동자 정규직 전환 촉구(NNG)	Negative
…	…	…	…

"문제는 한국 사회에서 시스템이 필요하다고 지시를 내릴 사람은 많은데 전통적으로 "노가다"를 뛸 사람은 없다는 겁니다. 이런 일은 남이 해야 하는 거라 생각하죠. 아니면 남이 했다가 자기한테 해가 되면 안 되니까 오만 가지 이유를 대서 이런 일은 하면 안 된다고 하고, 이런 일이 의료계에서만 있는 줄 알았어요. 사회 전반이 바뀌지 않으면 이 문제는 나아지지 않아요"라고 했다.

Manual Approach

- Sentiment Analysis Results

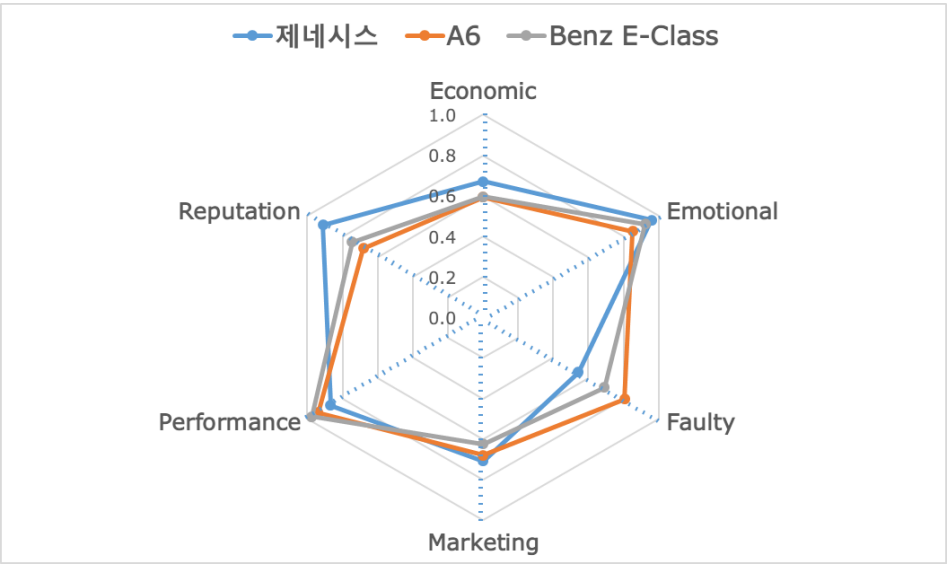




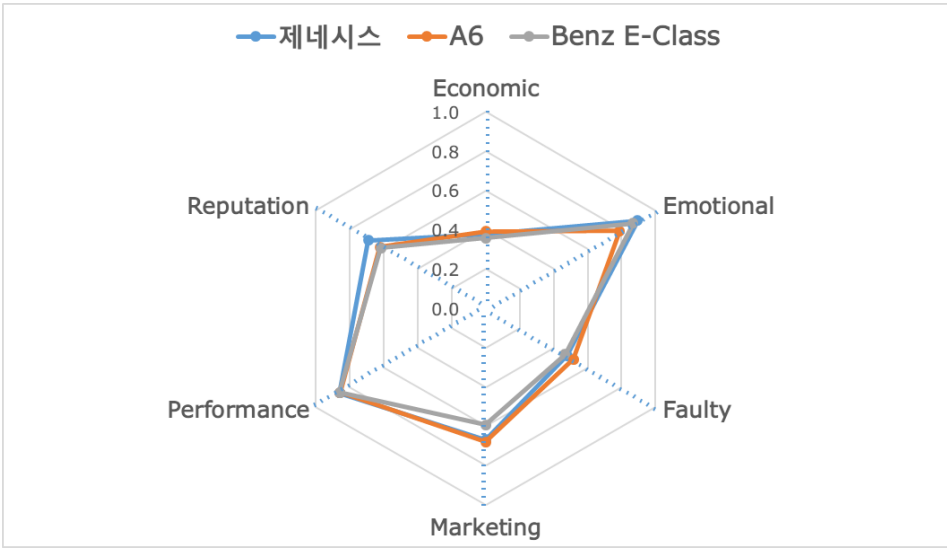
Manual Approach

- Sentiment Analysis Results

뉴스



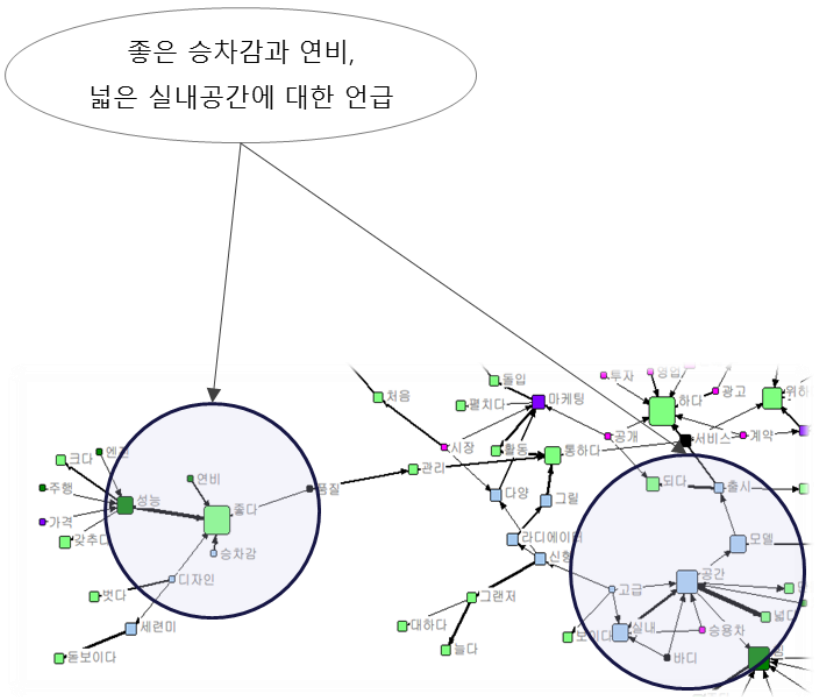
블로그



Manual Approach

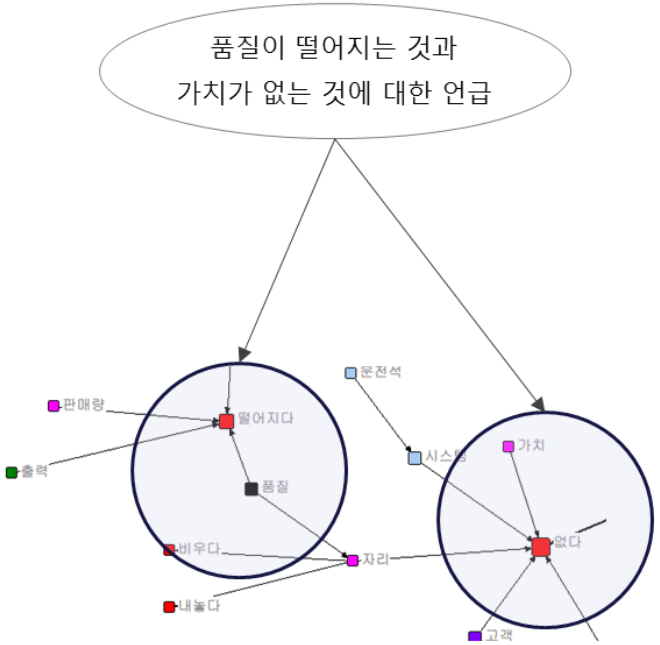
- Sentiment Analysis Results

긍정 연관어



긍정 연관어 네트워크 분석 결과 일부(뉴스)

부정 연관어



부정 연관어 네트워크 분석 결과 일부(뉴스)

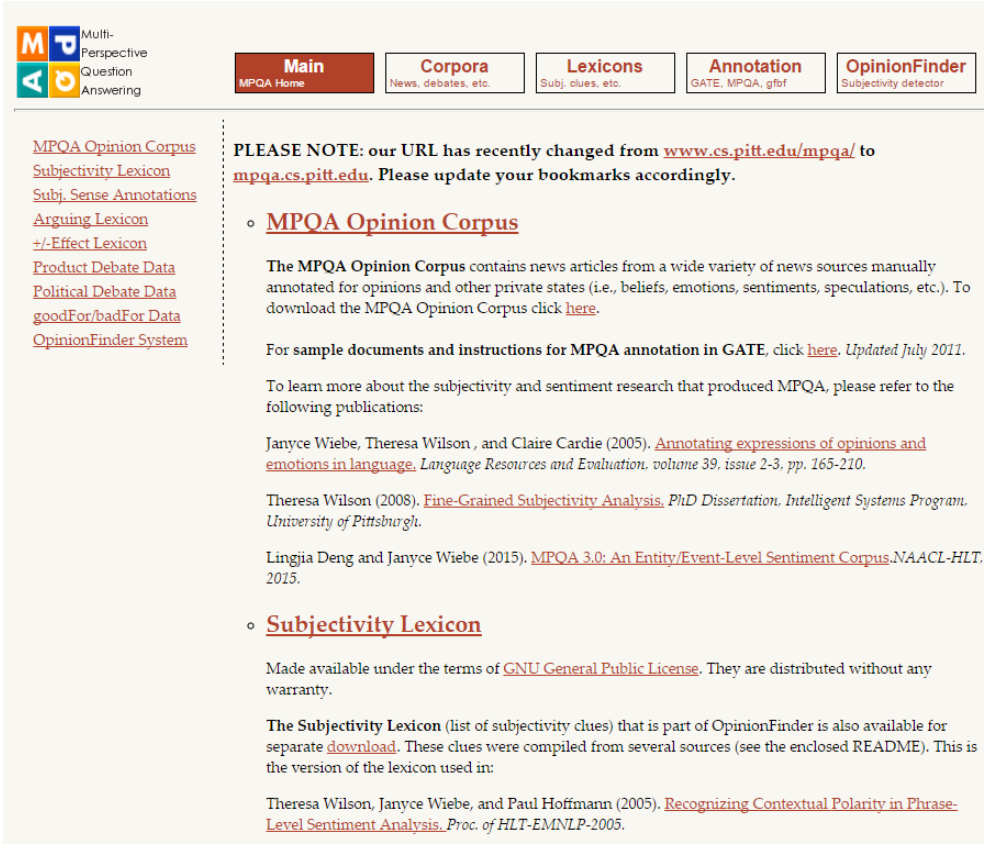
Dictionary- vs Corpus-based Approaches

Feldman (2013)

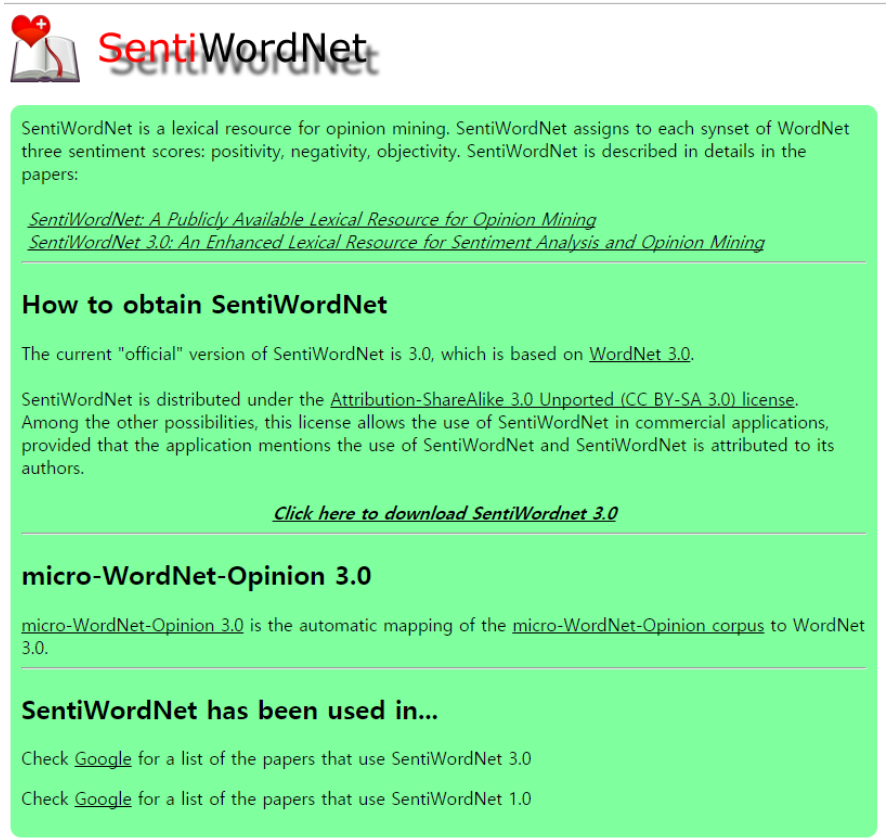
- Dictionary-based approaches
  - ✓ Typically use WordNet's synsets and hierarchies to acquire opinion words
    - Start with a small seed set of opinion words
  - ✓ Usually do not give domain or context dependent meanings
- Corpus-based approaches
  - ✓ Often use a double propagation between opinion words and the items they modify
    - Knowing an aspect can find the opinion word that modifies it
      - The **rooms** are ***spacious***
    - Knowing some opinion words can find more opinion words
      - The **rooms** are **spacious** and ***beautiful***
  - ✓ Can find domain dependent orientations
  - ✓ Require a large corpus to get good coverage

Dictionary-based Approaches

- Popular Sentiment Lexicon Database for English



http://mpqa.cs.pitt.edu/



http://sentiwordnet.isti.cnr.it/

Dictionary-based Approaches

- Popular Sentiment Lexicon Database for English

Opinion Lexicon (or Sentiment Lexicon)

- **Opinion Lexicon**: A list of English positive and negative opinion words or sentiment words (around 6800 words). This list was compiled over many years starting from our first paper (Hu and Liu, KDD-2004).
- **Comparative words**: A list of non-standard English compariative words and phrases for sentiment analysis. This list was compiled over many years starting from our first paper (Jindal and Liu, SIGIR-2006)
- Although necessary, having an opinion lexicon is far from sufficient for accurate sentiment analysis. See this paper: *Sentiment Analysis and Subjectivity* or the *Sentiment Analysis* book.
- Try Search for the Best Restaurant based on specific aspects, e.g., "best burger," "friendliest service." The system is a demo, which uses the lexicon (also phrases) and grammatical analysis for opinion mining.

Data Sets

- **Annotated: Customer Review Datasets (5 products)** associated with the paper (Hu and Liu, KDD-2004).
- **Annotated**: Additional Customer Review Datasets (9 products) some used in (Ding, Liu and Yu, WSDM-2008), which improves the lexicon-based method proposed in (Hu and Liu, KDD-2004)
- **Annotated**: More Customer Review Datasets (3 products) used in (Liu et al., IJCAI-2015)
- Amazon Product Review Data (more than 5.8 million reviews) used in (Jindal and Liu, WWW-2007, WSDM-2008; Lim et al, CIKM-2010; Jindal, Liu and Lim, CIKM-2010; Mukherjee et al. WWW-2011; Mukherjee, Liu and Glance, WWW-2012) for opinion spam (fake review) detection. You can also use it for sentiment analysis. It has information about reviewers, review texts, ratings, product info, etc. Due to the large file size, you may need to use *Download Accelerator Plus* (DAP) to download. If you use this data, please cite (Jindal and Liu, WSDM-2008).
- **Pros and cons dataset** used in (Ganapathibhotla and Liu, Coling-2008) for determining context (aspect) dependent sentiment words, which are then applied to sentiment analysis of comparative sentiences (comparative sentence dataset). The same form of Pros and Cons data was also used in (Liu, Hu and Cheng, WWW-2005).
- **Comparative sentence dataset** used in (Jindal and Liu, SIGIR-06) and (Jindal and Liu, AAAI-2006). **Comparative sentence dataset** used in (Ganapathibhotla and Liu, Coling-2008).
- **Blog author gender classification data set** associated with the paper (Mukherjee and Liu, EMNLP-2010)
- **Debate data set** used in (Mukherjee and Liu, ACL-2013; Mukherjee et al. ACL-2013).
- **Yelp Filtered Reviews** for Opinion spam or fake detection associated with the paper (Mukherjee et al.ICWSM-2013).

https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#lexicon

SenticNet

Helping machines to learn, leverage, love.

about sentic computing demo sentic api
downloads publications projects team

Talking about SenticNet is talking about concept-level sentiment analysis, that is, performing tasks such as polarity detection and emotion recognition by leveraging on semantics and linguistics in stead of solely relying on word co-occurrence frequencies.

In this context, SenticNet can be one of the following things:

1) a concept-level knowledge base;

2) a multi-disciplinary framework;

3) a private company.

As a ***knowledge base***, SenticNet provides a set of semantics, sentics, and polarity associated with 50,000 natural language concepts. In particular, semantics are concepts that are most semantically-related to the input concept (i.e., the five concepts that share more semantic features with the input concept), sentics are emotion categorization values expressed in terms of four affective dimensions (Pleasantness, Attention, Sensitivity, and Aptitude) and polarity is floating number between -1 and +1 (where -1 is extreme negativity and +1 is extreme positivity). The knoweldge base is downloadable for free as a standalone XML file and its latest version (released every two years) is also accessible as an API.

As a ***framework***, SenticNet consists of a set of tools and techniques for sentiment analysis combining commonsense reasoning, psychology, linguistics, and machine learning. In this context, SenticNet is more commonly referred to as *sentic computing*, a multi-disciplinary paradigm that goes beyond mere statistical approaches to sentiment analysis by focusing on a semantic-preserving representation of natural language concepts and on sentence structure.

As a ***company***, finally, SenticNet puts together the latest findings in concept-level sentiment analysis to offer easy-to-use state-of-the-art tools for big social data analysis that enable the automation of tasks such as brand positioning, trend discovery, and social media marketing in different modalities.

http://sentic.net/

Dictionary-based Approaches

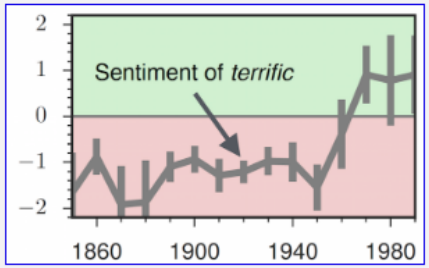
- SocialSent (Hamilton et al., 2016)

SocialSent: Domain-Specific Sentiment Lexicons for Computational Social Science

William L. Hamilton, Kevin Clark, Jure Leskovec, Dan Jurafsky

Introduction

The word *soft* may evoke positive connotations of warmth and cuddliness in many contexts, but calling a hockey player *soft* would be an insult. If you were to say something was *terrific* in the 1800s, this would probably imply that it was terrifying and awe-inspiring; today, *terrific* basically just implies that something is (pretty) good.



A word's sentiment or connotation depends on the domain or context in which it is used. However, previous computational work in natural language processing largely ignores this issue, and focuses and building and deploying generic domain-general sentiment lexicons.

SocialSent is a collection of code and datasets for performing *domain-specific* sentiment analysis. The SocialSent code package contains the SentProp algorithm for inducing domain-specific sentiment lexicons from unlabeled text, as well as a number of baseline algorithms.

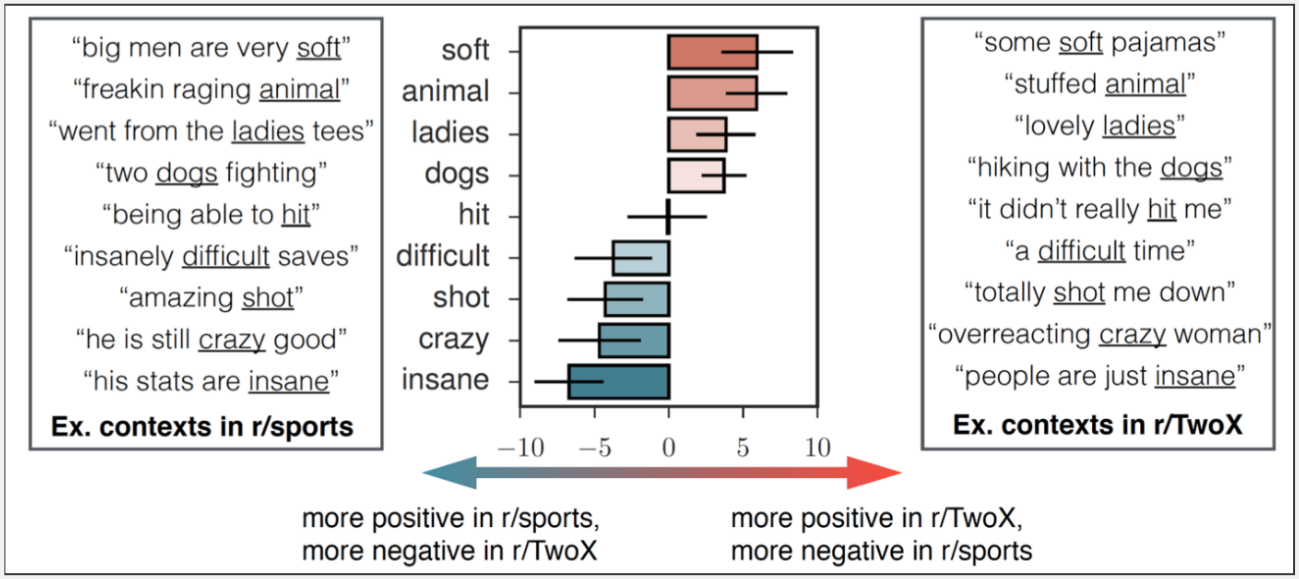
We have also released domain-specific historical sentiment lexicons for 150 years of English and community-specific sentiment lexicons for 250 "subreddit" communities from reddit.com. The historical lexicons reveal that more than 5% of sentiment-bearing words switched their polarity from 1850 to 2000, and the community-specific lexicons highlight how sentiment varies drastically between online communities.

The paper Inducing Domain-Specific Sentiment Lexicons from Unlabeled Corpora details the SentProp algorithm and describes the lexicons we induced.

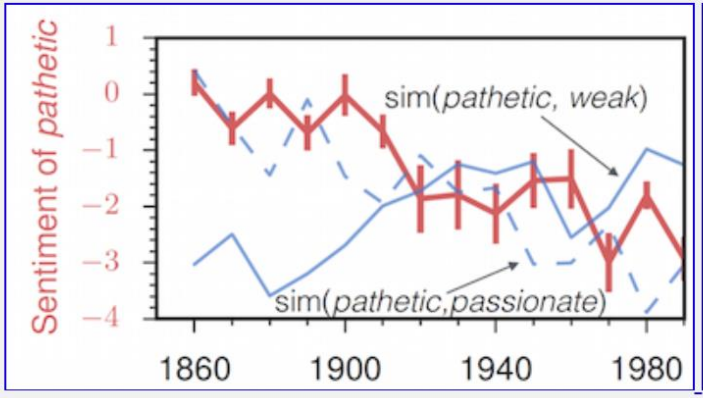
Dictionary-based Approaches

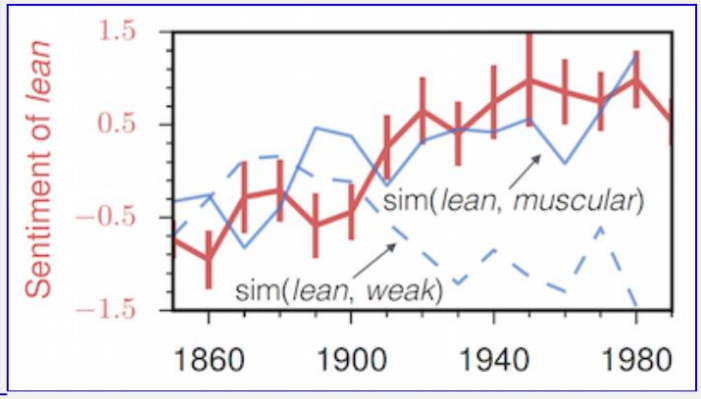
- SocialSent Highlights

1. **Word sentiment varies drastically between online communities**



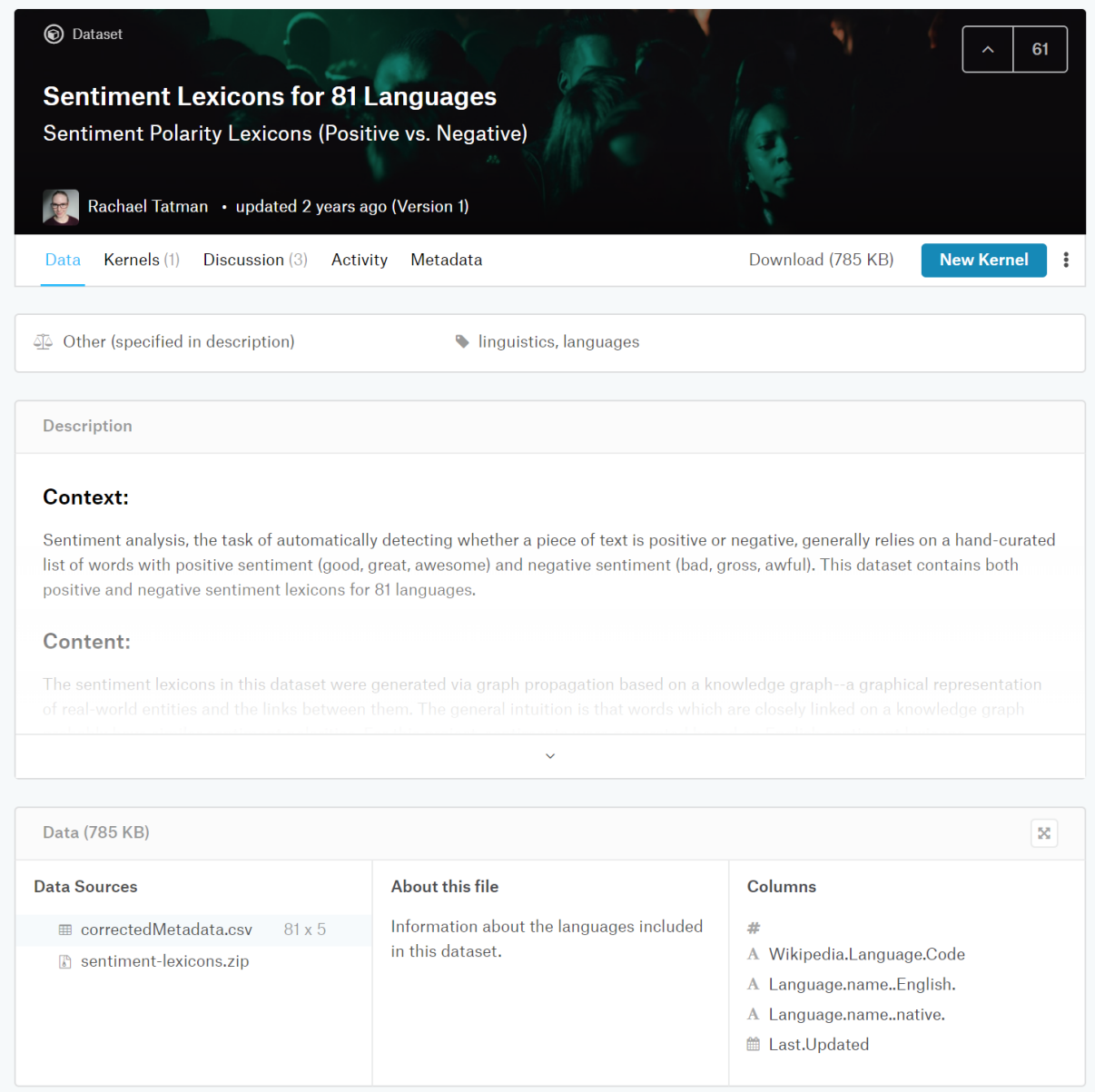
2. **Word sentiment varies dramatically over historical time-periods**





Dictionary-based Approaches

- Kaggle??



Dictionary-based Approaches

- Sentiment Lexicons by Kaggle
  - ✓883 positive lexicons, 1,235 negative lexicons for Korean

Positive lexicon examples

같은	정치	상	원하는	위로
더	자기	혁명	적절한	진정한
큰	고려	미	주요한	시인
주	쉽게	충분히	풍부한	일찍이
같이	최고의	아름다운	우수한	신성
매우	강력한	구조	전형적인	작업
잘	뛰어난	오른쪽	자유롭게	후한
볼	사람	상당한	특징	성공한
다양한	유명하다	지원	이해	고급
중요한	아주	단순한	장관	사랑
金	대	역사적	충분한	성공적으로
프로	개인	새롭게	참여	좋다
관련	자유	보호	비밀	유지
좋은	상당히	슈퍼	단일	수정
하다	초	지도	오른	보물
금	빠른	지속적으로	훌륭한	개혁
곧	기술	미리	종류	평화
시의	젊은	완전한	계약	힘든
유명한	right	정확히	막대한	변호사
비슷한	넓은	즉시	완전	사전
보다	인물	밝은	좋아하는	뜨거운
전체	특별한	전용	활발한	순수한
최고	德	크다	통일	굉장히
네	거대한	기능	꽤	지배
현대	정확한	방법	별의	강화
단	깊은	빨리	연	명예
사용할	양	진행	분명히	가능
강한	점	강하게	특유의	챔피언
선	원	간단한	승리	자치
작품	수상	해방	쉬운	성과

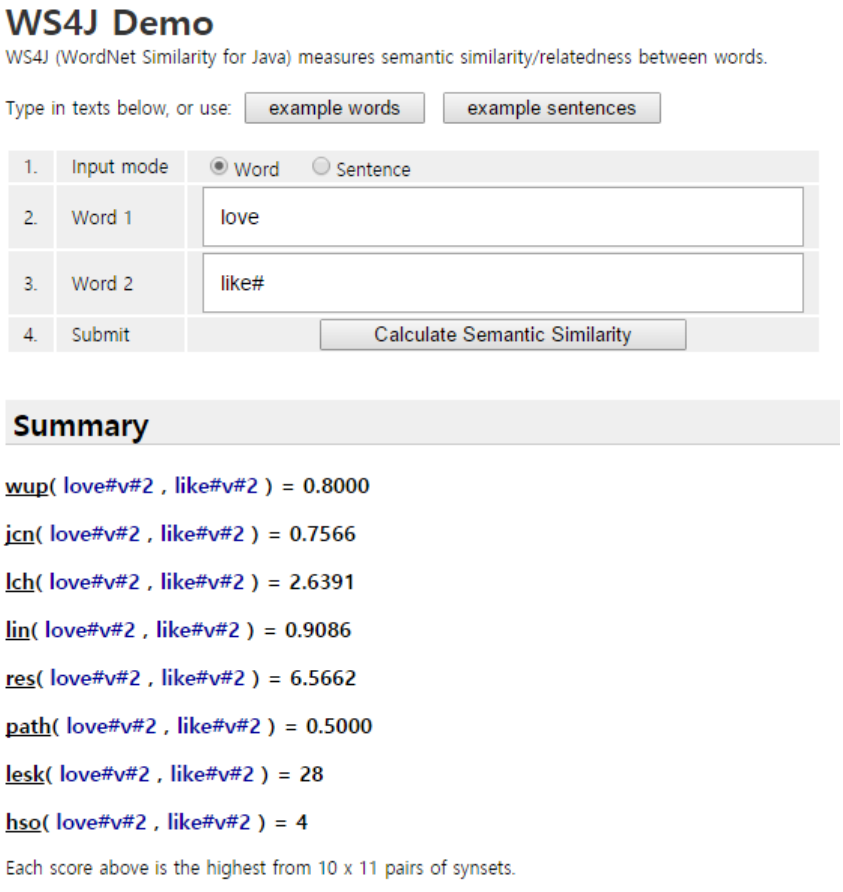
Negative lexicon examples

의	섬의	소수	지정	수상한
한	갑자기	시험	떨어지는	제외
다른	사망	어두운	문제의	힘
토론	독특한	전차	엄격한	포
특히	죽었다	결정	부족	지는
뒤	공격	죽을	분리	항구
없는	잘못된	비판	끊임없이	개인적으로
해	정의	상대	주장	폐지
없다	전쟁의	잃은	발견	테러
성	복잡한	무거운	분석	수사
크게	피해	향	타격	실패한
주의	부인	부족한	죽음	항의
만든	대상	나쁜	금지	고정
기타	적	길	암살	치열한
지방	잘못	기	김	분할
쓴	붙어	가난한	평범한	충돌
차단	아래의	화가	위험한	계승
긴	상태	색	줄거리	심하게
전혀	열	보지	소리	고문
삭제	강제	친일파	불법	위반
죽은	대신에	가을	분쟁	반란
문제	좁은	가상	연기	바이러스
사용하여	적의	노예	부하	정지
소설	뒤로	어렵다	붕괴	가사
신	발생	다리를	뚜렷한	가짜
본래	거친	경쟁	불가능한	부상
제작	약한	걸려	격렬한	살인
만화	칼	연결	범죄	망명
어려운	발표	회의	말기	짐
반대	지나치게	제한	잃었다	읽을

Dictionary-based Approaches

Feldman (2013)

- Dictionary-based Approach: Example
  - ✓ http://ws4jdemo.appspot.com/



WS4J Demo

WS4J (WordNet Similarity for Java) measures semantic similarity/relatedness between words.

Type in texts below, or use: example words | example sentences

1.	Input mode	Word / Sentence
2.	Word 1	love
3.	Word 2	like#
4.	Submit	Calculate Semantic Similarity

Summary

wup(love#v#2 , like#v#2) = 0.8000

jcn(love#v#2 , like#v#2) = 0.7566

lch(love#v#2 , like#v#2) = 2.6391

lin(love#v#2 , like#v#2) = 0.9086

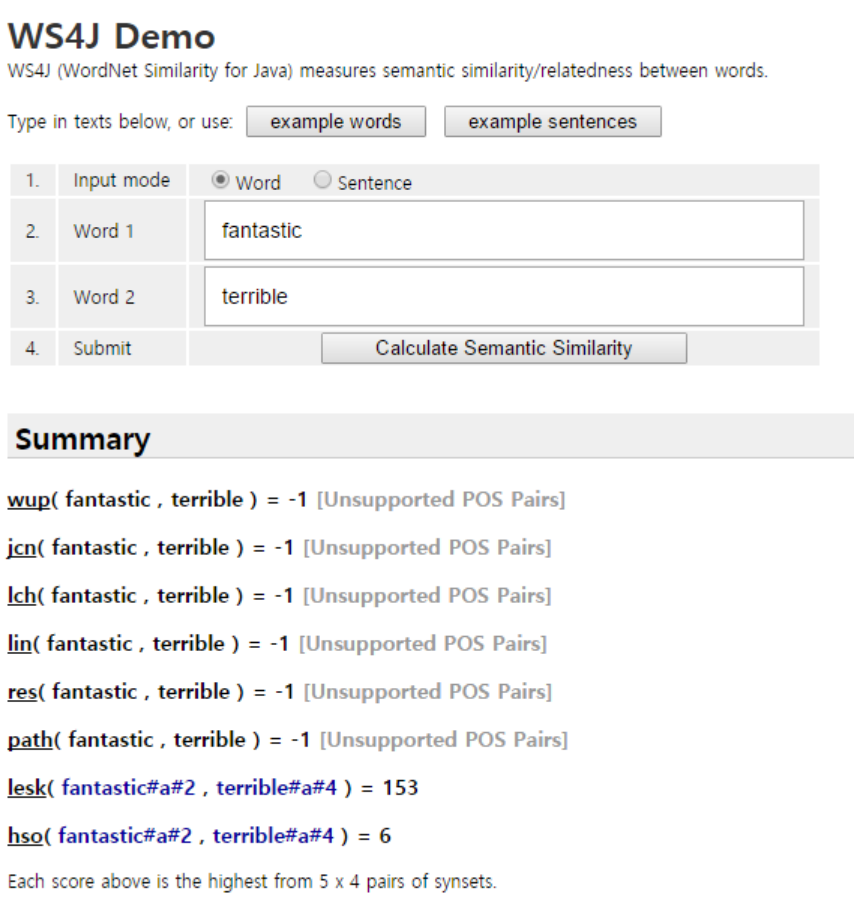
res(love#v#2 , like#v#2) = 6.5662

path(love#v#2 , like#v#2) = 0.5000

lesk(love#v#2 , like#v#2) = 28

hso(love#v#2 , like#v#2) = 4

Each score above is the highest from 10 x 11 pairs of synsets.



WS4J Demo

WS4J (WordNet Similarity for Java) measures semantic similarity/relatedness between words.

Type in texts below, or use: example words | example sentences

1.	Input mode	Word / Sentence
2.	Word 1	fantastic
3.	Word 2	terrible
4.	Submit	Calculate Semantic Similarity

Summary

wup(fantastic , terrible) = -1 [Unsupported POS Pairs]

jcn(fantastic , terrible) = -1 [Unsupported POS Pairs]

lch(fantastic , terrible) = -1 [Unsupported POS Pairs]

lin(fantastic , terrible) = -1 [Unsupported POS Pairs]

res(fantastic , terrible) = -1 [Unsupported POS Pairs]

path(fantastic , terrible) = -1 [Unsupported POS Pairs]

lesk(fantastic#a#2 , terrible#a#4) = 153

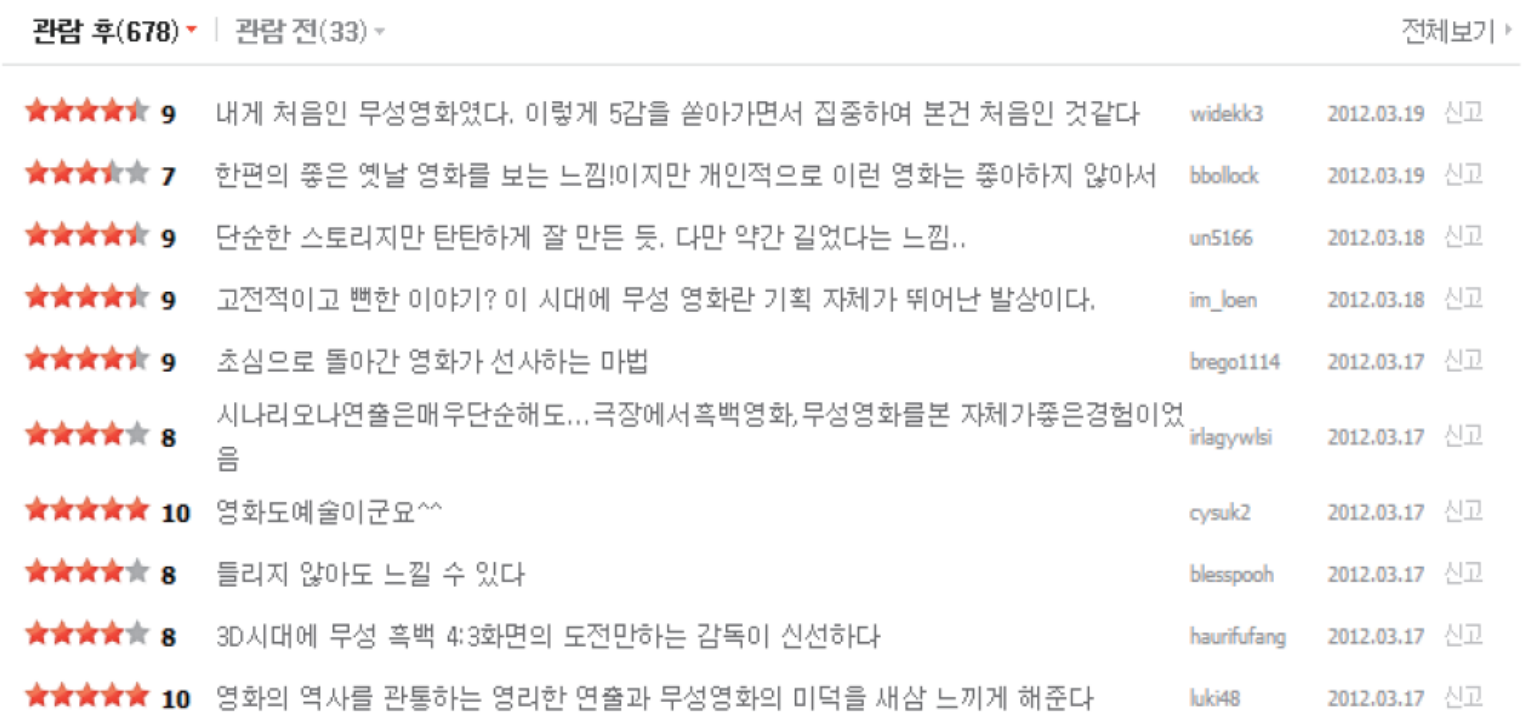
hso(fantastic#a#2 , terrible#a#4) = 6

Each score above is the highest from 5 x 4 pairs of synsets.

Corpus-based Approaches

Hur et al. (2016)

- Building Up a Sentiment Dictionary based on Movie Review Data



관람 후(678) | 관람 전(33) 전체보기

9	내게 처음인 무성영화였다. 이렇게 5감을 쏟아가면서 집중하여 본건 처음인 것같다	widekk3	2012.03.19	신고
7	한편의 좋은 옛날 영화를 보는 느낌!이지만 개인적으로 이런 영화는 좋아하지 않아서	bbollock	2012.03.19	신고
9	단순한 스토리지만 탄탄하게 잘 만든 듯. 다만 약간 길었다는 느낌..	un5166	2012.03.18	신고
9	고전적이고 뻔한 이야기? 이 시대에 무성 영화란 기획 자체가 뛰어난 발상이다.	im_loen	2012.03.18	신고
9	초심으로 돌아간 영화가 선사하는 마법	brego1114	2012.03.17	신고
8	시나리오나연출은매우단순해도...극장에서흑백영화,무성영화를본 자체가좋은경험이었음	irlagywlsi	2012.03.17	신고
10	영화도예술이군요^^	cysuk2	2012.03.17	신고
8	들리지 않아도 느낄 수 있다	blesspooh	2012.03.17	신고
8	3D시대에 무성 흑백 4:3화면의 도전만하는 감독이 신선하다	haurifufang	2012.03.17	신고
10	영화의 역사를 관통하는 영리한 연출과 무성영화의 미덕을 새삼 느끼게 해준다	luki48	2012.03.17	신고

Corpus-based Approaches

Hur et al. (2016)

- Building Up a Sentiment Dictionary based on Movie Review Data
  - ✓ T-test for testing the difference of review ratings with and without a word

$$R(r_{i,j}, w_q) = \begin{cases} 0 & \text{if } w_q \notin r_{i,j} \\ r(r_{i,j}) & \text{if } w_q \in r_{i,j} \end{cases}$$

$$Score(w_q) = E(w_q) = \frac{1}{n(w_q)} \sum_{i=1}^{m} \left(\sum_{j=1}^{n_i} R(r_{i,j}, w_q) \right)$$

$$Var(w_q) = \frac{1}{n(w_q) - 1} \sum_{i=1}^{m} \left(\sum_{j=1}^{n_i} \left(R(r_{i,j}, w_q) - Score(w_q) \right)^2 \right)$$

Corpus-based Approaches

Hur et al. (2016)

- Building Up a Sentiment Dictionary based on Movie Review Data
  - ✓ T-test for testing the difference of review ratings with and without a word

$$\text{Testing} : T_w = \frac{E(W) - E(w)}{\sqrt{\frac{s_W^2}{n(W)} + \frac{s_w^2}{n(w)}}} \quad \begin{cases} \text{Positive if} & T_w > t_{(\alpha;v)} \\ \text{Negative if} & T_w < -t_{(\alpha;v)} \\ \text{Neutral} & Otherwise \end{cases}$$

$$\text{where} \quad v = \frac{\left(s_W^2/n(W) + s_w^2/n(w)\right)^2}{\frac{\left(s_W^2/n(W)\right)^2}{n(W)-1} + \frac{(s_w^2/n(w))^2}{n(w)-1}}$$

Corpus-based Approaches

Hur et al. (2016)

- Building Up a Sentiment Dictionary based on Movie Review Data
  - ✓ Example

Collocation	Avg	Stdev.	Count	NMovies	T	t-dist	Diag
멋지/vv	9.52	1.22	1,617	341	61.88	1.96	Positive
실망/NNG_시키/XSV_지/ECD_않/VXV	9.68	0.97	829	260	60.87	1.96	Positive
재밌/VA_게/ECD_보/VXV	9.13	1.70	4,679	565	59.90	1.96	Positive
재미/NNG_잇/VV	9.57	1.45	2,004	379	59.74	1.96	Positive
깔끔/XR_하/XSA	9.24	1.22	1,336	358	48.02	1.96	Positive
잘/MAG_만들/VV_ㄴ/ETD_영화/NNG	9.26	1.49	1,903	342	47.47	1.96	Positive
못/MAG_보시/VV_는/ETD_분/NNG	7.83	2.37	12	9	0.29	2.2	Neutral
남자/NNG_둘/NNG	7.78	3.18	36	34	0.26	2.03	Neutral
영웅/NNG_이야기/NNG	7.71	2.76	48	24	0.18	2.01	Neutral
복수극/NNG	7.68	2.75	104	38	0.17	1.98	Neutral
코믹/NNG_성/XSN	7.7	2.83	44	31	0.16	2.02	Neutral
도저히/MAG_모르/VV	4.27	3.28	15	15	-3.99	2.14	Negative
손발/NNG_오글거리/VV	4.41	3.8	22	18	-3.99	2.08	Negative
무리수/NNG	5.62	3.11	39	25	-4.06	2.02	Negative
안타깝/VA_ㄴ/ETD_영화/NNG	6.42	3.21	119	84	-4.13	1.98	Negative
좀/MAG_지겹˙VA	6.62	2.2	89	72	-4.38	1.99	Negative
억지/NNG_이/VCP	6.3	2.96	121	88	-4.98	1.98	Negative

AGENDA

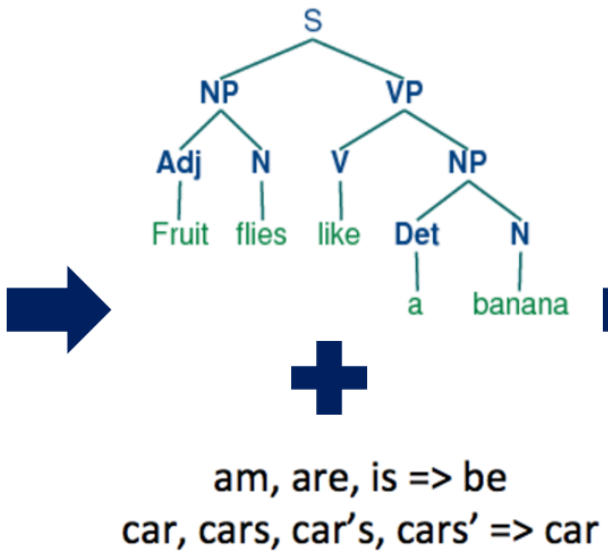
Machine Learning-based Approach

조수현 외 (2016)

- Regression-based Approach

  - TripAdvisor 사이트에서 서울 여행지에 대한 평가 댓 글과 5점 척도의 평점 수집(21620개)
  - 형태소 분석(POS tagging)을 통해 형용사, 동사, 명사 추출
  - Lemmatization을 실시하여 문서 단어 행렬(document term matrix) 구성
  - 문서 단어 행렬을 사전에 정의한 서울의 관광지 유형에 따라 분류

Crawling을 통한 데이터 수집



POS tagging + Lemmatization

유형	문서 수	장소
명소&랜드마크	14,443	남산, 경복궁, 인사동 등
쇼핑&시장	3,589	명동쇼핑거리, 코엑스, 광장시장 등
박물관&전시관	3,588	전쟁기념관, 국립민속박물관 등

유형 별 관광지, 문서의 수

	Term1	Term2	...	Term(m)
Doc1	1	0	...	7
Doc2	0	0	...	0
...	...	...	...	...
Doc(n)	2	0	...	2

Document term matrix

Machine Learning-based Approach

조수현 외 (2016)

- Regression-based Approach

  - 유형별 관광지의 document term matrix에서 각 유형별로 벌점화 회귀분석을 활용하여 감정사전 구축
  - 10-fold cross validation을 적용하여 엘라스틱 넷 학습
  - Validation data에 대해 여러 지표를 이용하여 성능을 비교한 결과 감정사전으로서 타당한 성능확보

유형 별 관광지의 감정사전 워드 클라우드

	명소&랜드마크	쇼핑&시장	박물관&전시관
어휘	2,826	1,319	1,329
Elastic net	747	249	242

유형 별 관광지의 감정사전 어휘개수

	명소&랜드마크	쇼핑&시장	박물관&전시관
ACC	0.8246	0.7465	0.8717
REC	0.8911	0.8028	0.9451
PRE	0.8999	0.8686	0.9149
BCR	0.6444	0.6506	0.4561
MAE	0.5727	0.6446	0.5625
COR	0.4434	0.3887	0.3575

유형 별 관광지의 감정사전 성능

Machine Learning-based Approach

조수현 외 (2016)

- Regression-based Approach
  - ✓ Elastic Net
    - Response variable: rating of a review
    - Explanatory variables: words used in the review
    - Use the regression coefficient as the sentiment score of the corresponding word

$$\hat{\beta} = \arg\min_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda_2\|\beta\|^2 + \lambda_1\|\beta\|_1$$

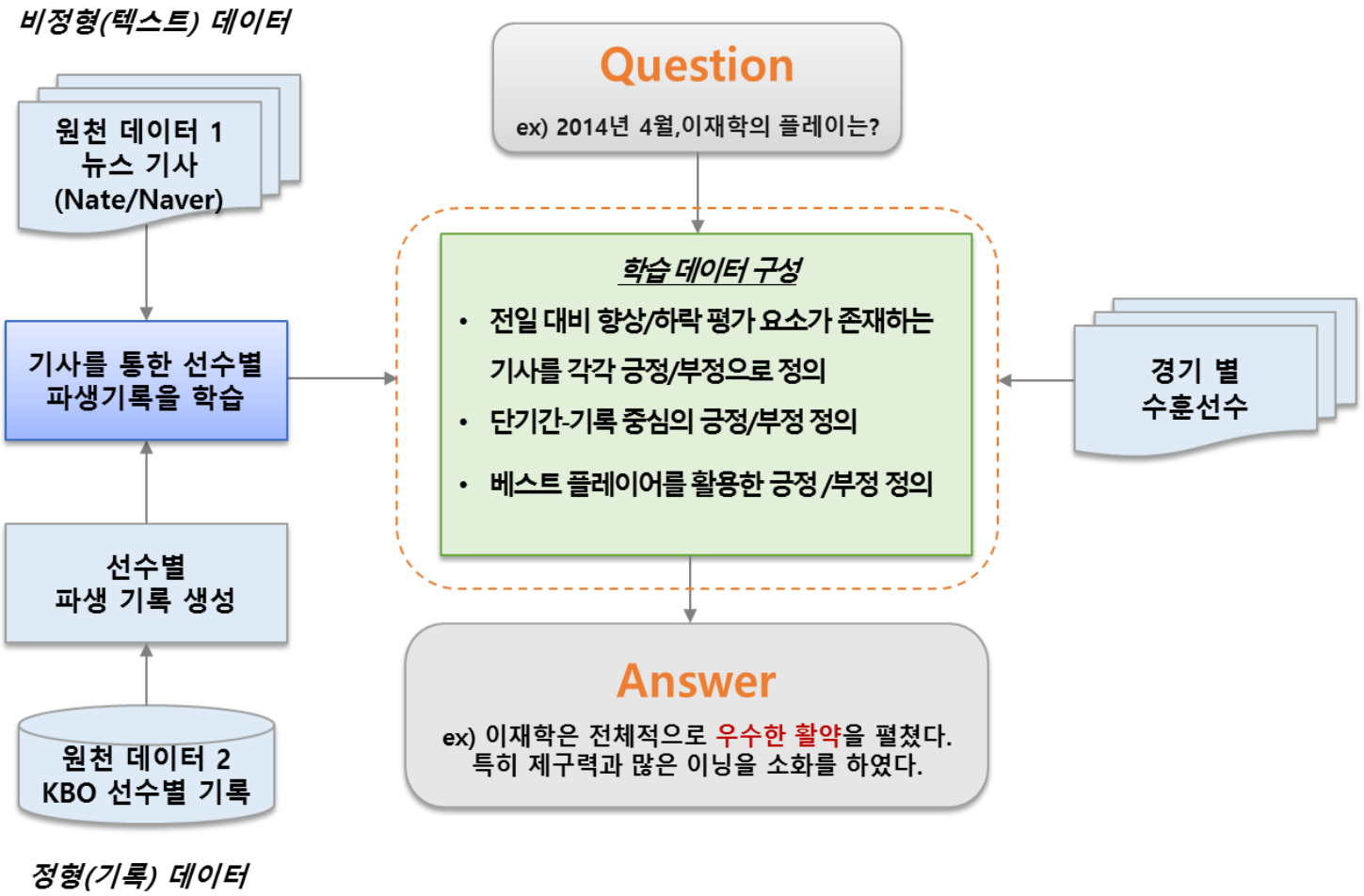
- The ℓ_1 part of the penalty generates a sparse model.
- The quadratic part of the penalty
  - Removes the limitation on the number of selected variables;
  - Encourages *grouping effect*;
  - Stabilizes the ℓ_1 regularization path.

http://web.stanford.edu/~hastie/TALKS/enet_talk.pdf

Machine Learning-based Approach

Park et al. (2016+)

- Neural Network-based Approach
  - ✓ Evaluating a player based on the sentiment of news articles



Machine Learning-based Approach

Park et al. (2016+)

- Neural Network-based Approach
  - ✓ Evaluating a player based on the sentiment of news articles

Best Player Case

<2014.05.29 NC 테임즈 관련기사>

날짜	선수	타수	안타	타점	득점	타율	홈런	볼넷	삼진
05-28	테임즈	3	2	0	2	0.315	0	1	0
볼넷 비율	삼진 비율	볼/삼 비율	출루율	장타율	ops	ISO	타석당 홈런	RC	BABIP
0.33	0	Inf	0.75	1.00	1.75	0.68	0	2.25	0.67

↓

날짜	선수	**타수**	**안타**	**타점**	**득점**	**타율**	**홈런**	볼넷	삼진
05-29	테임즈	**6**	**5**	**7**	**4**	**0.333**	**2**	0	0
볼넷 비율	삼진 비율	볼/삼 비율	**출루율**	**장타율**	**ops**	**ISO**	**타석당 홈런**	**RC**	**BABIP**
0	0	NA	**0.83**	**2.00**	**2.83**	**1.67**	**0.333**	**10.00**	**0.75**

↓

타수	안타	타점	득점	타율	홈런	타율	홈런	출루율	장타율	ops	ISO	타석당 홈런	RC	BABIP
+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1

볼넷	삼진	볼넷	볼넷 비율	삼진 비율	볼/삼 비율
0	0	0	0	0	0

↓

Best player 1

Non-best Player Case

<2014.05.29 NC 손시헌 관련기사>

이전 경기기록

날짜	선수	타수	안타	타점	득점	타율	홈런	볼넷	삼진
05-28	손시헌	4	1	0	0	0.292	0	0	0
볼넷 비율	삼진 비율	볼/삼 비율	출루율	장타율	ops	ISO	타석당 홈런	RC	BABIP
0	0	0	0.25	0.25	0.5	-0.042	0	0.25	0.25

↓

당일 경기기록

날짜	선수	타수	안타	타점	**득점**	타율	홈런	**볼넷**	삼진
05-29	손시헌	4	1	0	**1**	0.291	0	**1**	0
볼넷 비율	삼진 비율	**볼/삼 비율**	**출루율**	장타율	**ops**	**ISO**	**타석당 홈런**	**RC**	**BABIP**
0.25	0	**Inf**	**0.4**	0.25	**0.65**	**-0.041**	**0**	**0.40**	**0.25**

↓

Target

득점	볼넷	볼넷 비율	출루율	ops	ISO	타석당 홈런	RC	BABIP
+1	+1	+1	+1	+1	+1	+1	+1	+1

타수	안타	타점	타율	홈런	삼진	삼진 비율	장타율
0	0	0	0	0	0	0	0

↓

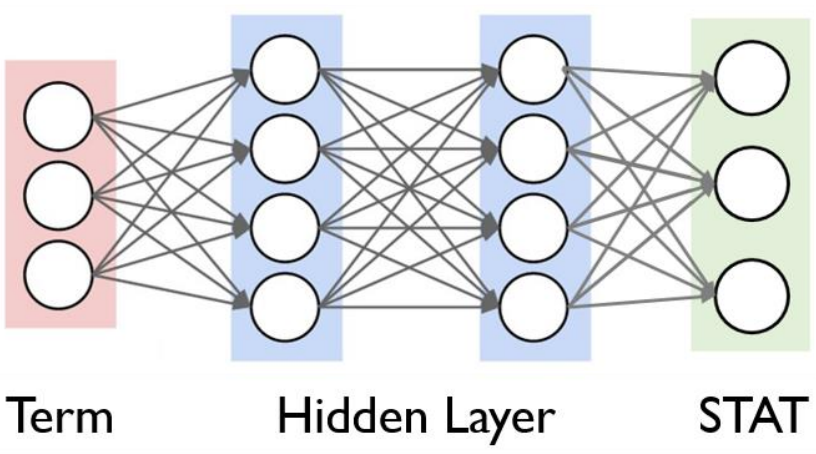
Best player 0

Machine Learning-based Approach

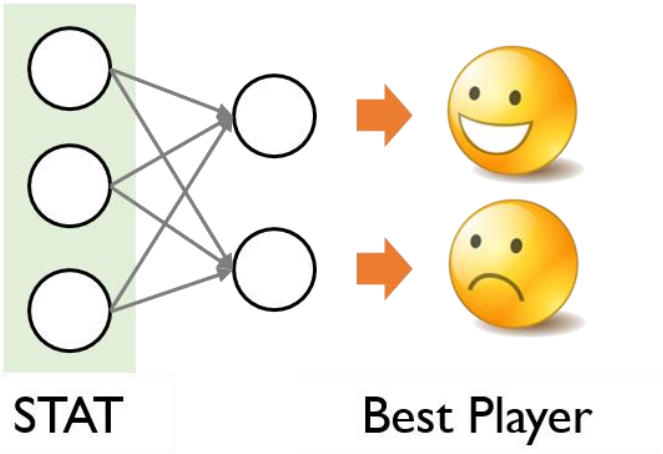
Park et al. (2016+)

- Neural Network-based Approach
  - ✓ Evaluating a player based on the sentiment of news articles

<문장을 벡터화하여 기록의 변화를 학습하는 모델>



<기록 변화를 통해 우수 선수를 학습하는 모델>



투수의 경우

9,670 terms 9670 : 4000 : 2000 : 1000 : 18 STAT 18개 18 : 2 우수 / 비우수

타자의 경우

8,686 terms 8686 : 4000 : 2000 : 1000 : 9 STAT 9개 9 : 2 우수 / 비우수

Machine Learning-based Approach

Park et al. (2016+)

- Neural Network-based Approach
  - ✓ Evaluating a player based on the sentiment of news articles

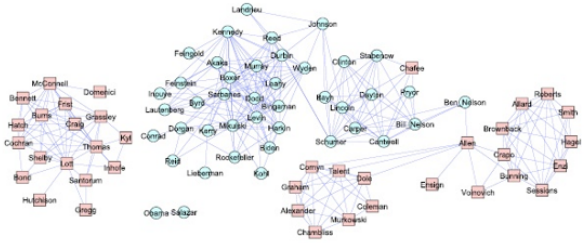
Player	Line	Score	승	패배	피안타	피홈런	볼넷	실점	평균 자책점	이닝당 삼진	이닝당 볼넷	삼진 /볼넷	방어율
웨버	NC 웨버는 7이닝 동안 1안타 3볼넷 10탈삼진으로 무실점 호투했다.	0.991	1	0	0	0	1	0	0	1	1	1	0
이재학	이재학 '10승', 3년만에 달라진 위상 증명하다	0.986	1	0	0	1	0	1	0	0	0	1	0
이태양	'이태양 완벽투' NC, LG 상대로 창단 첫 스윕	0.966	1	0	0	0	0	0	0	0	0	1	0
에릭	한편, 이날 7이닝 3실점으로 잘 던지며 7경기 만에 국내 첫 승을 거둔 에릭은 "굉장히 흥분된다.	0.803	1	0	1	0	0	0	0	0	0	0	0
찰리	어이없는 실책 2개가 나오자 찰리는 흔들리기 시작했다.	0.506	0	1	1	1	0	1	0	1	0	1	1
찰리	찰리가 대량 실점을 했지만 그 과정에서 3루수 모창민과 1루수 테임즈의 실책이 동반되면서 자책은 1점 밖에 되지 않았다.	0.506	0	1	1	1	0	1	0	1	0	1	1
원종현	NC 원종현이 패전투수가 됐다.	0.293	0	1	0	0	0	1	1	1	0	1	1
원종현	세 번째 투수로 등판한 원종현 역시 1사 후 8번 김성현에게 솔로포를 맞았다.	0.102	0	0	0	1	1	0	1	1	1	0	0
이재학	NC 선발 이재학은 8이닝 8피안타(2홈런) 5탈삼진 3볼넷 2실점을 기록했다.	0.079	0	0	1	1	0	1	1	0	0	0	1
이혜천	이혜천이 올라온 뒤 한꺼번에 5점을 내주면서 맥빠진 경기가 되고 말았다.	0.066	0	0	1	1	0	1	0	0	1	0	0

Machine Learning-based Approach

서덕성 외 (2016)

- Semi-supervised Learning-based Approach

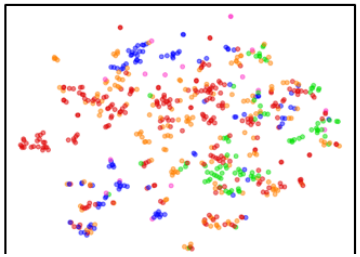
❖ Framework



데이터 준비 | Word2Vec | Construct network | Graph based SSL | 성능 평가

01 Text + scaled score 데이터 수집

02 고차원으로 표현된 단어를 저차원으로 임베딩



03 임베딩 공간에서 거리기반 네트워크 구축

04 명백한 label을 미리 선정/할당 label propagation을 이용 score 추정

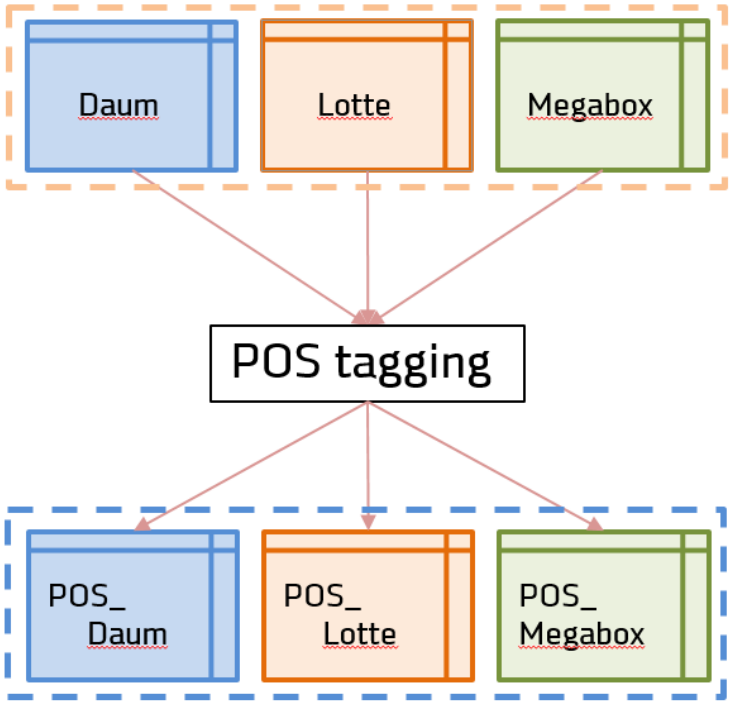
Term	Score
재미있다	1
...	...
지루하다	-1

05 추정한 감성 점수 성능 평가

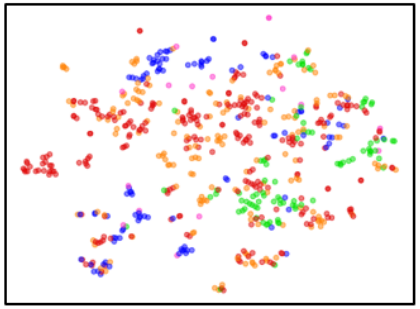
Machine Learning-based Approach

서덕성 외 (2016)

- Semi-supervised Learning-based Approach



Term	d_1	d_2	...	d_n
명작	0.234	0.891	...	0.963
아쉽다	0.477	0.443	...	0.564
...	...	...	...	...
연기력	0.119	0.863	...	0.787



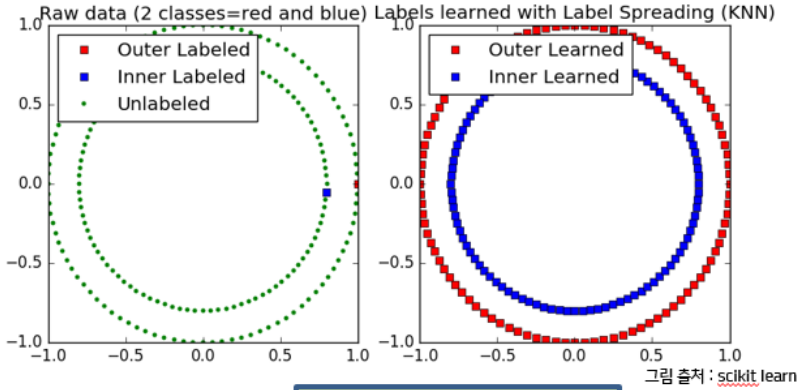
Machine Learning-based Approach

서덕성 외 (2016)

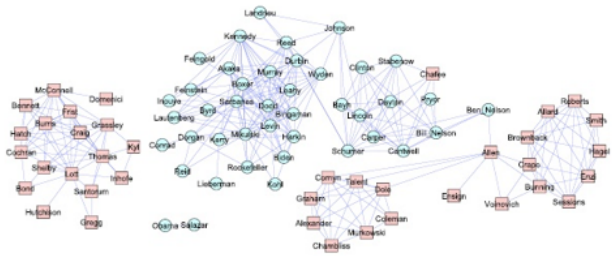
- Semi-supervised Learning-based Approach

데이터 준비 → Word2Vec → Construct network → Graph based SSL → 성능 평가

Term	d_1	d_2	...	d_n
명작	0.234	0.891	...	0.963
아쉽다	0.477	0.443	...	0.564
...	...	...	...	...
연기력	0.119	0.863	...	0.787



Distance Matrix



중심성

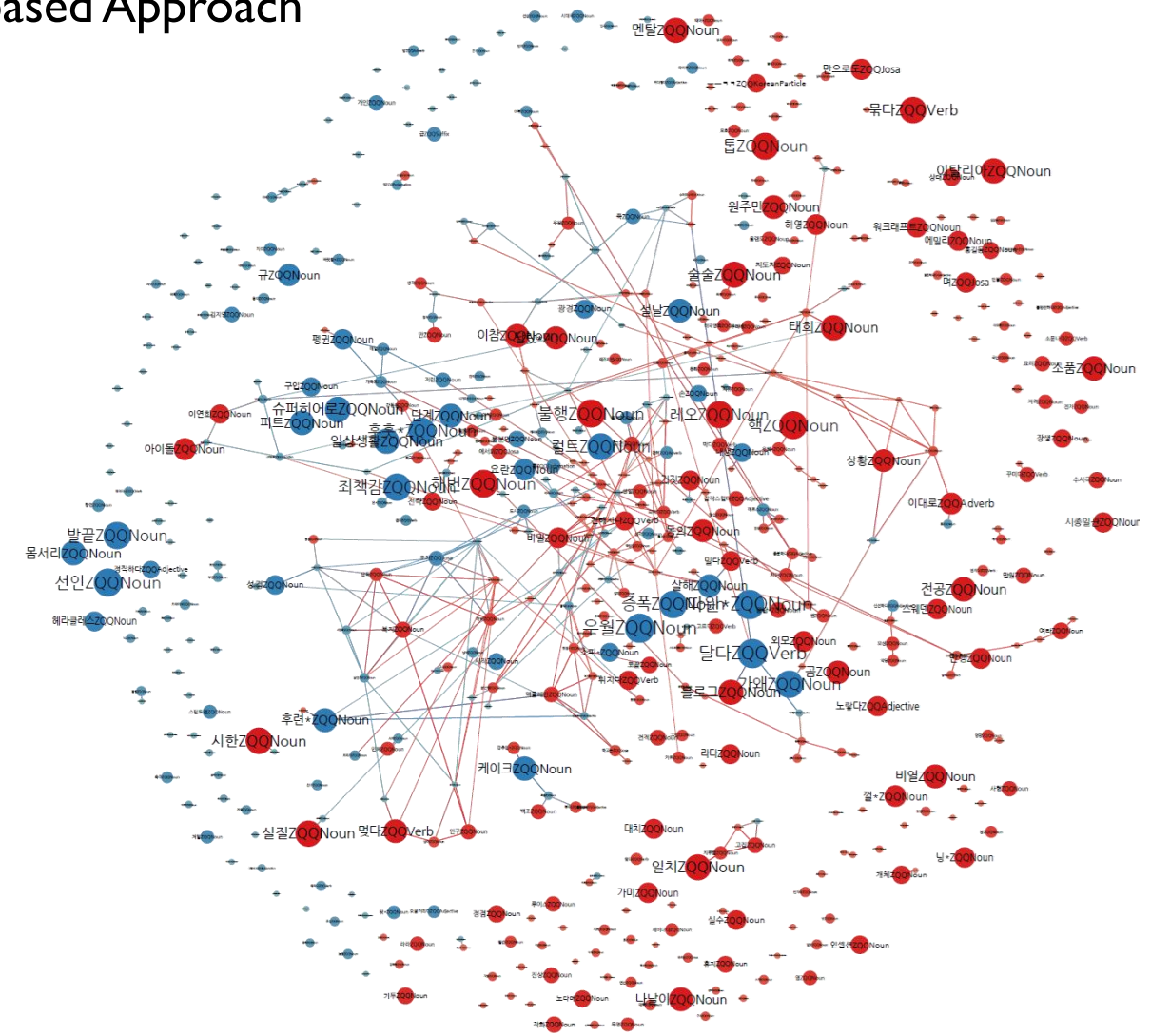
Pre-labeled word

상위 50	상위 50
긍정 (331)	부정 (219)

Machine Learning-based Approach

서덕성 외 (2016)

- Semi-supervised Learning-based Approach



Machine Learning-based Approach

서덕성 외 (2016)

- Semi-supervised Learning-based Approach
  - ✓ Pre-labeled words

Positive pre-labeled words			Negative pre-labeled words		
좋다	역시	연기력	억지	지루	때우다
훌륭하다	즐겁다	아름답다	아쉬움	약하다	부담
재밌다	기대하다	긴장감	절대	흠	부끄럽다
사랑	좋아하다	매력	짜증	진부	고통
재미있다	멋지다	짱	지루함	싫다	난해
감동	대박	빠지다	밉다	별루	거슬리다
기대	웃다	멋있다	이상하다	심하다	망치다
최고	눈물	울다	킬링타임	식상하다	애매
재미	추천	코믹	필요없다	까다	거지
괜찮다	슬프다	오랜만	어이	졸리다	질질

Machine Learning-based Approach

서덕성 외 (2016)

- Semi-supervised Learning-based Approach
  - ✓ Sentiment scores for unlabeled words

Positive		Negative	
Words	Sentiment Score	Words	Sentiment Score
슬펏	1	딸리	-0.9576
먹먹	1	별루였어	-0.5458
이뿌	0.8282	겉도	-0.4673
괞찮	0.5956	드럽	-0.4140
잼잼	0.5126	부족	-0.3713
신나요	0.4357	심해	-0.3449
눈시울	0.4089	어설퍼	-0.3123
짱	0.4002	미흡	-0.3062
왈칵	0.4001	지루	-0.2728
존잘	0.3851	빈약	-0.2531

Machine Learning-based Approach

서덕성 외 (2016)

- Semi-supervised Learning-based Approach

✓ Performance evaluation (Word sentiment)

	ACC	REC	PRE	F1	BCR
Proposed method	0.9311	0.9288	0.9596	0.9439	0.9319
Lasso regression	0.4909	0.3807	0.6269	0.4737	0.5003

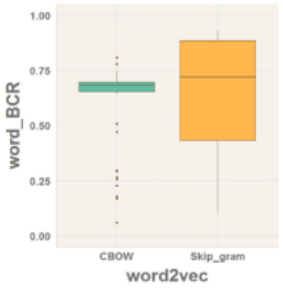
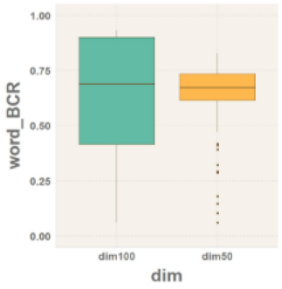
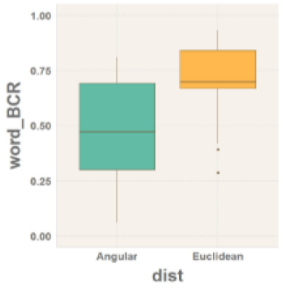
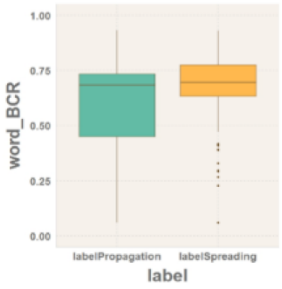
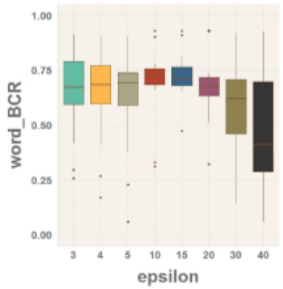
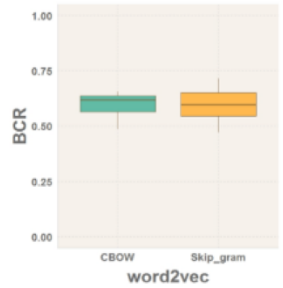
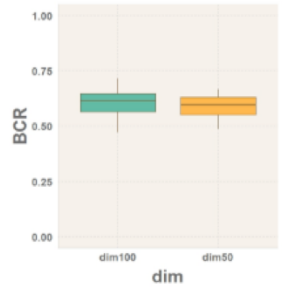
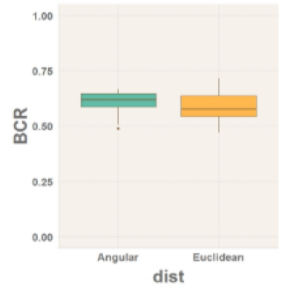
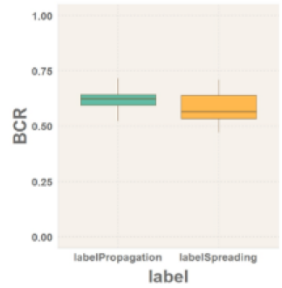
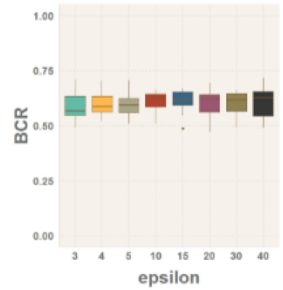
✓ Performance evaluation (Review polarity)

	ACC	REC	PRE	F1	BCR
Proposed method	0.7832	0.8163	0.9113	0.8612	0.7162
Lasso regression	0.7045	0.7704	0.8778	0.8206	0.4234

Machine Learning-based Approach

서덕성 외 (2016)

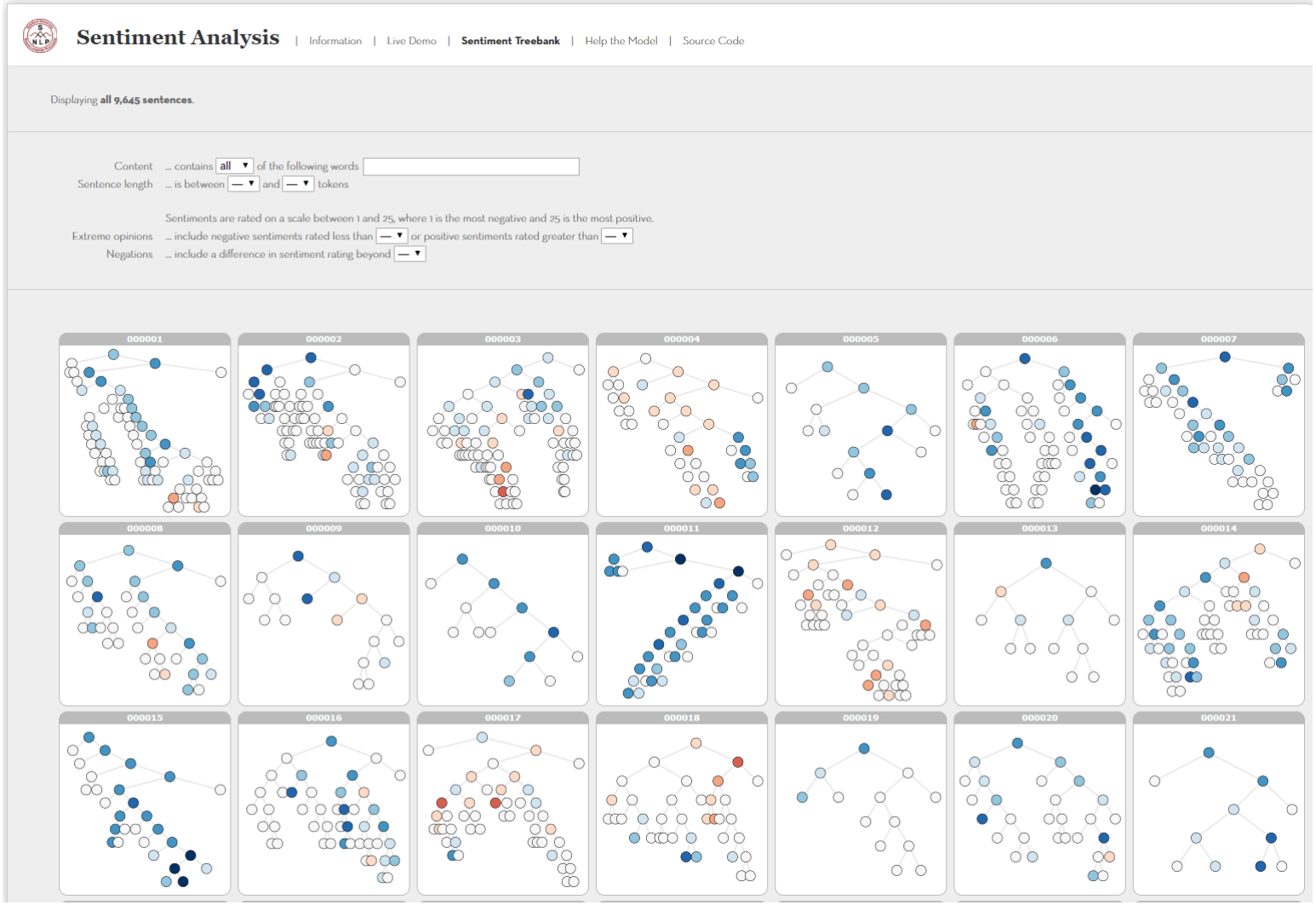
- Semi-supervised Learning-based Approach
  - ✓ Parameter sensitivity

BCR	Word2Vec	Embedding dimension	Distance measure	Graph based SSL	ϵ (percentile of distance)
Word sentiment					
Movie score					

Machine Learning-based Approach

Socher et al. (2013)

- Recursive Neural Tensor Network (RNTN)

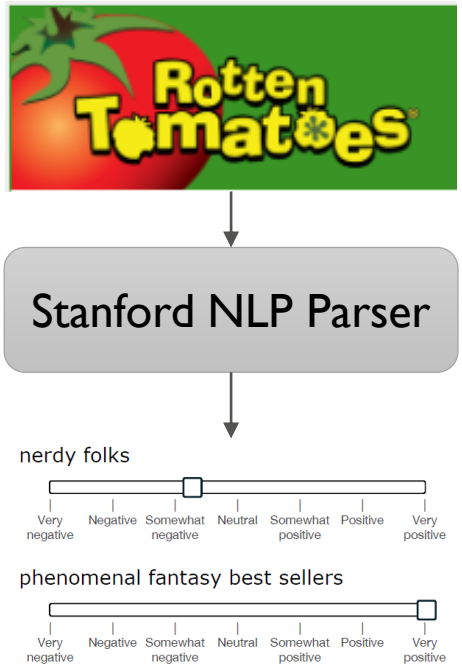


https://nlp.stanford.edu/sentiment/treebank.html

Machine Learning-based Approach

Socher et al. (2013)

- Recursive Neural Tensor Network (RNTN): Sentiment Treebank
  - ✓ Fully labeled parse trees that allows for a complete analysis of the compositional effects of sentiment in languages

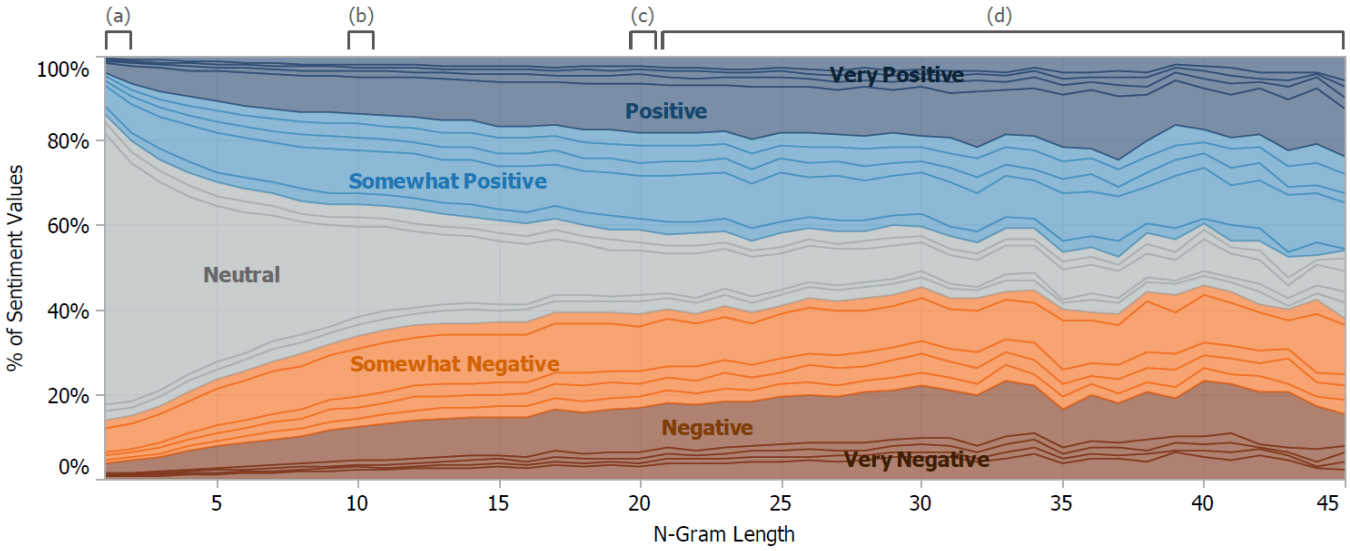


- 10,662 sentences
- 1,100 cases are split into multiple sentences
- 11,855 sentences with 215,154 phrases
- Label all phrases using Amazon Mechanical Turk
- Slider has 25 different values
- Each phrase is annotated by 3 human judges

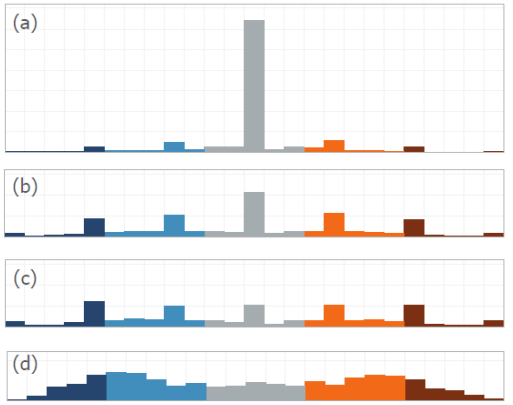
Machine Learning-based Approach

Socher et al. (2013)

- Recursive Neural Tensor Network (RNTN): Sentiment Treebank



Distributions of sentiment values for (a) unigrams, (b) 10-grams, (c) 20-grams, and (d) full sentences.



✓ Majority of shorter phrases are neutral

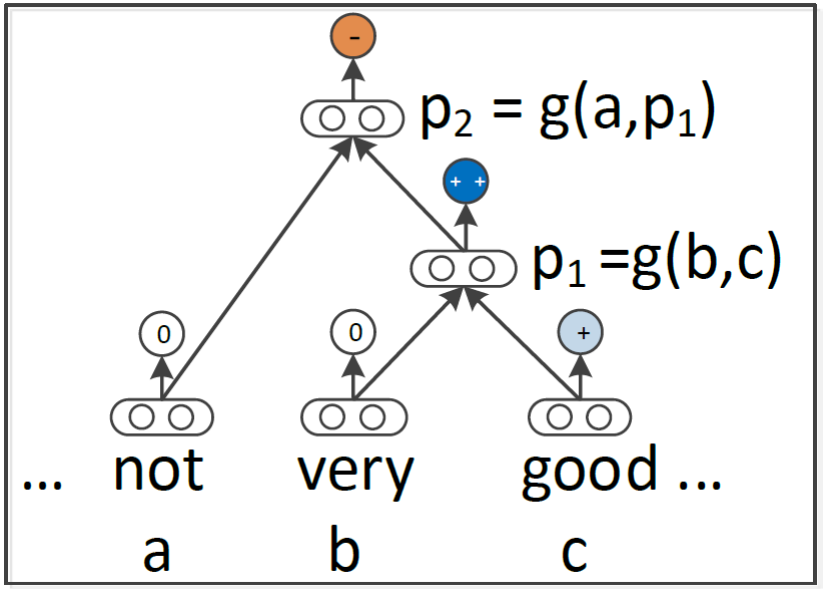
✓ Stronger sentiment often builds up in longer phrases

✓ Main experiment is to recover the five labels for phrases of all lengths

Machine Learning-based Approach

Socher et al. (2013)

- Recursive Neural Tensor Network (RNTN)
  - ✓ Benchmark Model 1: Recursive Neural Models



- Each word is represented as a d-dimensional vector
- Posterior probability over labels given the word vector:

$$y^{word} = softmax(W_s \cdot word)$$

Machine Learning-based Approach

Socher et al. (2013)

- Recursive Neural Tensor Network (RNTN)
  - ✓ Benchmark Model 1: Recursive Neural Models

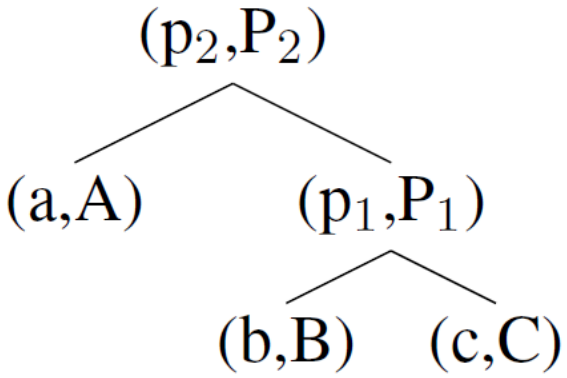
$$p_1 = f\left(W \begin{bmatrix} b \\ c \end{bmatrix}\right), \quad p_2 = f\left(W \begin{bmatrix} a \\ p_1 \end{bmatrix}\right)$$

- $f = tanh$ is a standard element-wise nonlinearity
- Parent vectors must be of the same dimensionality to be recursively compatible and be used as input to the next composition
- Each parent vector is given to the same softmax classifier to computer its label probabilities

Machine Learning-based Approach

Socher et al. (2013)

- Recursive Neural Tensor Network (RNTN)
  - ✓ Benchmark Model 2: Matrix-Vector RNN (MV-RNN)
    - Limitations of the standard RNN: Most of the parameters are associated with words and each composition function that computes vectors for longer phrases depends on the actual words being combined



$$p_1 = f\left(W \begin{bmatrix} Cb \\ Bc \end{bmatrix}\right), \quad P_1 = f\left(W_M \begin{bmatrix} B \\ C \end{bmatrix}\right)$$

Machine Learning-based Approach

Socher et al. (2013)

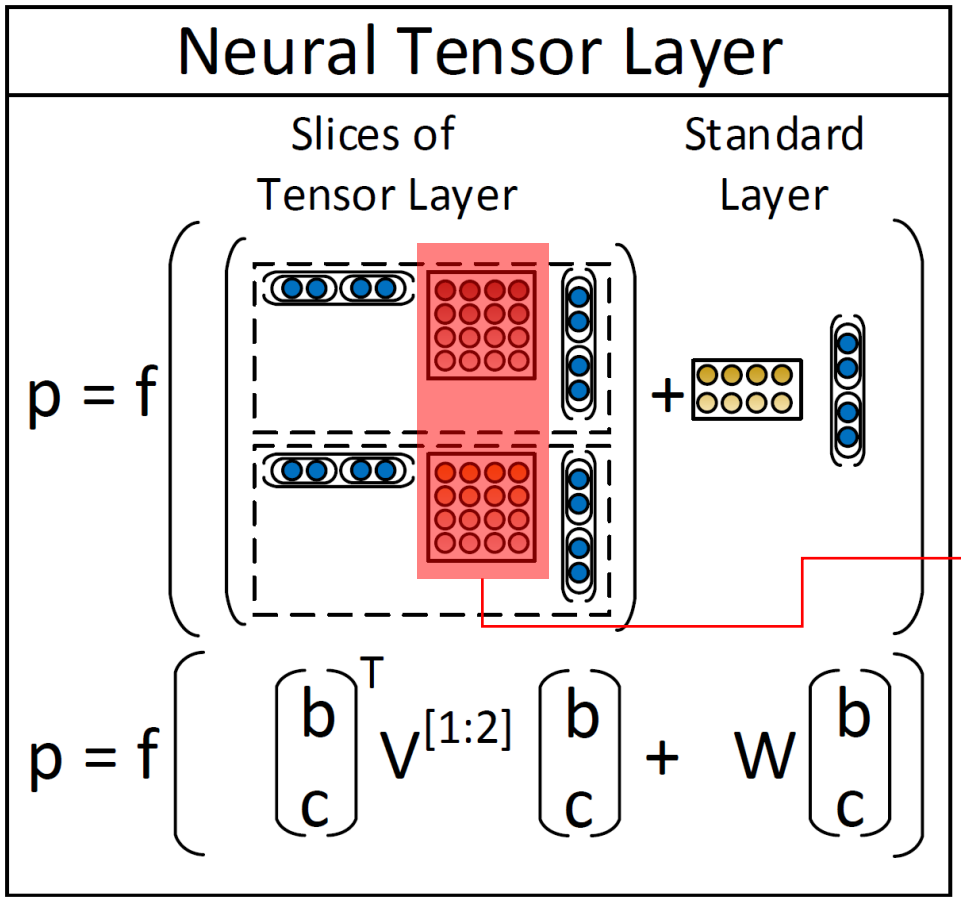
- Recursive Neural Tensor Network (RNTN)
  - ✓ Limitation of the MV-RNN: the number of parameters becomes very large and depends on the size of the vocabulary

	Standard RNN	MV-RNN	RNTN
Word vector	$d \times \|V\|$	$d \times \|V\|$	$d \times \|V\|$
Softmax	$C \times d$	$C \times d$	$C \times d$
Composition	$d \times 2d$	$W : d \times 2d$ $W_M : d \times 2d$	$W : d \times 2d$ $V^{[1:d]} : 2d \times 2d \times d$

Machine Learning-based Approach

Socher et al. (2013)

- Recursive Neural Tensor Network (RNTN)



$$p_1 = f\left(\begin{bmatrix} b \\ c \end{bmatrix}^T V^{[1:d]} \begin{bmatrix} b \\ c \end{bmatrix} + W \begin{bmatrix} b \\ c \end{bmatrix} \right)$$

$$p_2 = f\left(\begin{bmatrix} a \\ p_1 \end{bmatrix}^T V^{[1:d]} \begin{bmatrix} a \\ p_1 \end{bmatrix} + W \begin{bmatrix} a \\ p_1 \end{bmatrix} \right)$$

Each slice of the tensor can capture a specific type of composition

$$E(\theta) = \sum_i \sum_j t_j^i \log y_j^i + \lambda ||\theta||^2$$

Machine Learning-based Approach

Socher et al. (2013)

- Recursive Neural Tensor Network (RNTN)
  - ✓ Experiments

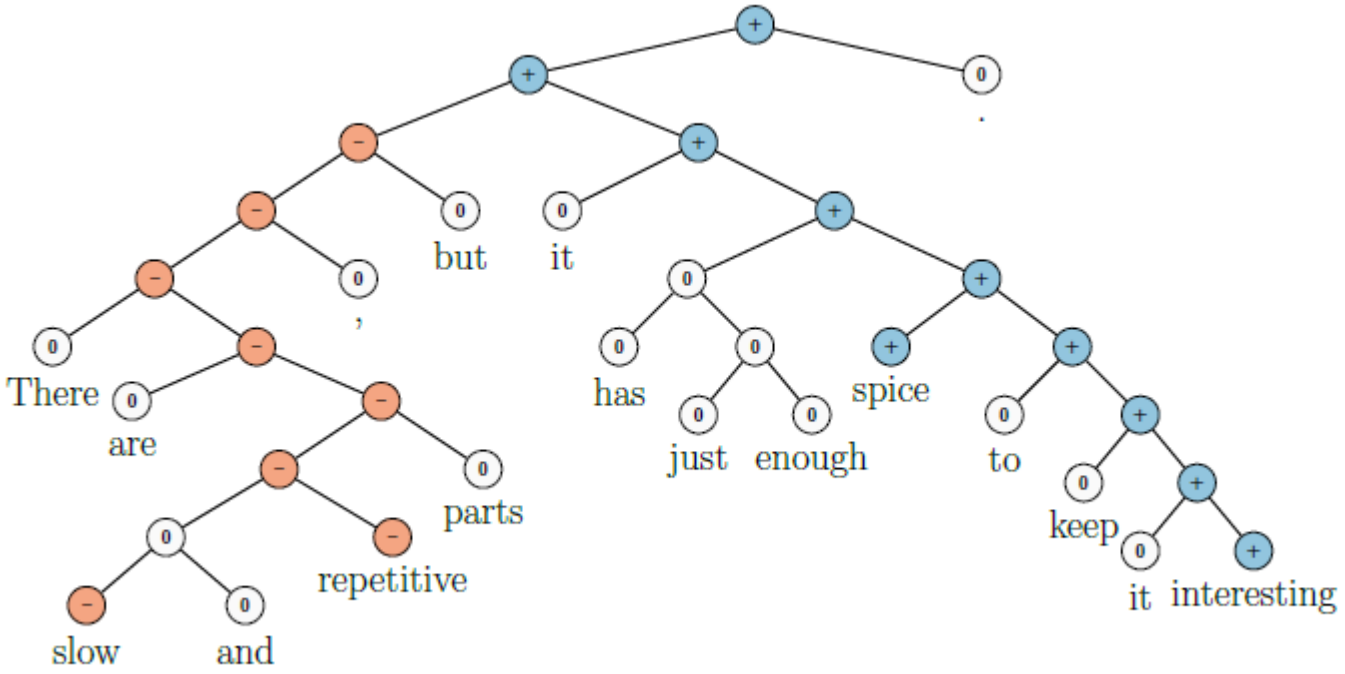
Model	Fine-grained		Positive/Negative	
	All	Root	All	Root
NB	67.2	41.0	82.6	81.8
SVM	64.3	40.7	84.6	79.4
BiNB	71.0	41.9	82.7	83.1
VecAvg	73.3	32.7	85.1	80.1
RNN	79.0	43.2	86.1	82.4
MV-RNN	78.7	44.4	86.8	82.9
RNTN	**80.7**	**45.7**	**87.6**	**85.4**

    - Optimization performance for all models was achieved at word vector sizes between 25 and 35

Machine Learning-based Approach

Socher et al. (2013)

- Recursive Neural Tensor Network (RNTN)
  - ✓ Model Analysis: Contrastive Conjunction: X but Y (131 cases)
    - RNTN: 41%, MV-RNN: 37%, RNNL 36%, biNB: 27%

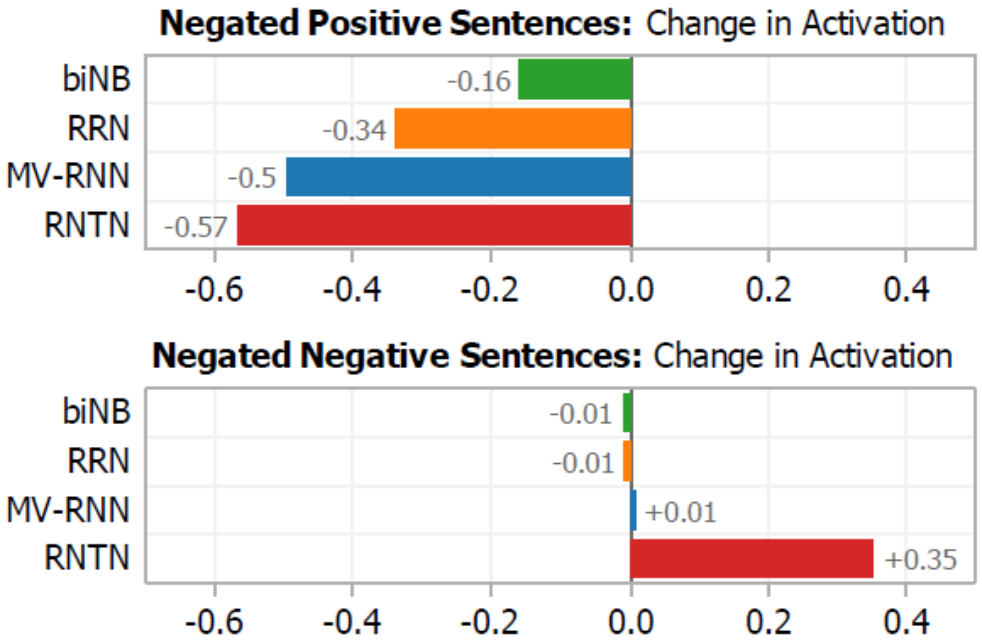


Machine Learning-based Approach

Socher et al. (2013)

- Recursive Neural Tensor Network (RNTN)
  - ✓ Model Analysis: High Level Negation

Model	Accuracy	
	Negated Positive	Negated Negative
biNB	19.0	27.3
RNN	33.3	45.5
MV-RNN	52.4	54.6
RNTN	**71.4**	**81.8**



Machine Learning-based Approach

Socher et al. (2013)

- Recursive Neural Tensor Network (RNTN)
  - ✓ Model Analysis: Most positive and negative phrases

n	Most positive n-grams	Most negative n-grams
1	engaging; best; powerful; love; beautiful	bad; dull; boring; fails; worst; stupid; painfully
2	excellent performances; A masterpiece; masterful film; wonderful movie; marvelous performances	worst movie; very bad; shapeless mess; worst thing; instantly forgettable; complete failure
3	an amazing performance; wonderful all-ages triumph; a wonderful movie; most visually stunning	for worst movie; A lousy movie; a complete failure; most painfully marginal; very bad sign
5	nicely acted and beautifully shot; gorgeous imagery, effective performances; the best of the year; a terrific American sports movie; refreshingly honest and ultimately touching	silliest and most incoherent movie; completely crass and forgettable movie; just another bad movie. A cumbersome and cliche-ridden movie; a humorless, disjointed mess
8	one of the best films of the year; A love for films shines through each frame; created a masterful piece of artistry right here; A masterful film from a master filmmaker,	A trashy, exploitative, thoroughly unpleasant experience ; this sloppy drama is an empty vessel.; quickly drags on becoming boring and predictable.; be the worst special-effects creation of the year


ANY
questions?

References

Serrano-Guerrero et al. (2015)

Research Papers

- 조수현, 김보섭, 박민식, 이기창, 강필성*. (2017). 여행 사이트 리뷰를 활용한 관광지 만족도 요인 추출 및 평가. 대한산업공학회지 43(1), 62-71.
- 서덕성, 모경현, 박재선, 이기창, 강필성. (2016). Word sentiment score evaluation based on semi supervised learning in a distributed representation, 2016 대한산업공학회 추계학술대회, 고려대학교, 서울, 11월 19일.
- Hamilton, W. L., Clark, K., Leskovec, J., & Jurafsky, D. (2016, November). Inducing domain-specific sentiment lexicons from unlabeled corpora. In Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing (Vol. 2016, p. 595). NIH Public Access.
- Hur, M., Kang, P., & Cho, S. (2016). Box-office forecasting based on sentiments of movie reviews and Independent subspace method. Information Sciences, 372, 608-624.
- Feldman, R. (2013). Techniques and applications for sentiment analysis. Communications of the ACM, 56(4), 82-89.
- Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends® in Information Retrieval, 2(1–2), 1-135.
- Park, Y., Kim, H., Lee, H., Kim, D., Kim, S., and Kang, P. (2016+). A deep learning-based sports player evaluation model for question-answering system based on game statistics and news articles, under review.
- Serrano-Guerrero, J., Olivas, J. A., Romero, F. P., & Herrera-Viedma, E. (2015). Sentiment analysis: a review and comparative analysis of web services. Information Sciences, 311, 18-38.
- Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y., & Potts, C. (2013, October). Recursive deep models for semantic compositionality over a sentiment treebank. In Proceedings of the conference on empirical methods in natural language processing (EMNLP) (Vol. 1631, p. 1642).

References

Research Papers (cont')

- Zhang, L., Wang, S., & Liu, B. (2018). Deep learning for sentiment analysis: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8(4), e1253.

Other Materials

- Feldman, R. (2013). Sentiment Analysis Tutorial. The 23rd International Joint Conference on Artificial Intelligence (IJCAI'13), Beijing, China. (http://ijcai13.org/files/tutorial_slides/tf4.pdf)
- Maynard, D., Funk, A., Bontcheva, K. (2012). Opinion Mining: Exploiting the sentiment of the Crowd. Tutorial at University of Sheffield, UK.
- Moghaddam S. & Ester, M. (2013). Opinion Mining in Online Reviews: Recent Trends. Tutorial at WWW 2013.