Sentimentanalyse for norsk tekst

SANT @ NRK

Erik Velldal og Lilja Øvrelid

Institutt for informatikk, Universitetet i Oslo

29. november 2017

Language Technology Group (LTG)

- Språkteknologigruppa, Institutt for Informatikk, UiO.
- 12 ansatte (4 faste).
- Lingvistikk + maskinlæring.
- Kjært barn: language engineering, computational linguistics, language technology, natural language processing / NLP.
- Foreløpig mye som mangler for norsk,
- inkludert ressurser for sentimentanalyse (SA).

Hva er SA?

- Aka *opinion mining*.
- Identifisere subjektive meninger som uttrykkes i tekst.
- Forskjellig granularitet:
- Positiv / negativ?
- Ulike klasser av følelser.
- Dokument-nivå vs setnings- og frase-nivå.
- Eksempler på bruksområder: mediaovervåking, markedsanalyse, 'produktsporing', opinion- og trendanalyse, og mye mer.

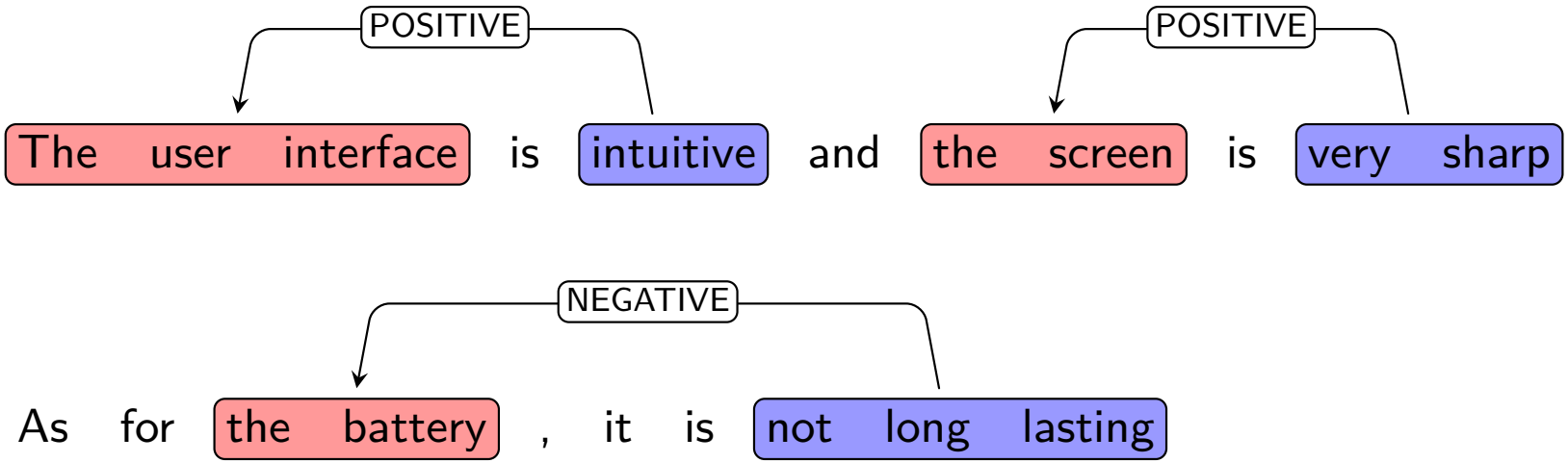
SANT

- Sentiment Analysis for Norwegian Text
- Samarbeid mellom LTG/IFI, Schibsted, Aller og NRK.
- Har søkt NFR/IKTPLUSS om en PhD + en postdok.
- Foreløpig fått invilget midler til en halvårig pilot.

Hva vi skal snakke om

- Hva er målet?
- Hva har vi gjort i fase 1? (jun.–nov.)
- Hva er planen videre; fase 2? (2018 – 2021)

Finkornet SA



- Veiledet maskinlæring: krever at vi har eksempler å lære fra.
- Vi må altså manuelt annotere tekst for å lage treningsdata.
- Noe av det vi har søkt om NFR-midler for.

Terningkast

- I det halvårige prøveprosjektet (fase 1):
- Begynner med mer grovkornet SA; dokument-nivå.
- Ide: terningkast for anmeldelser som eksempler.
- Representerer grad av positivitet/negativitet.
- Tren modeller til å predikere terningkast for en ny tekst.

Fase 1

- To vit.ass.'er: Eivind og Cathrine.
- Norwegian Review Corpus
- NoReC.
- https://github.com/ltgoslo/norec
- Publikasjon sendt til fagfellevurdering.
- I gang med å trene baseline-modeller for å predikere terningkast.

Source	# Reviews
VG	11,888
Dagbladet	5,305
Stavanger Aftenblad	5,146
P3.no	5,017
DinSide.no	2,944
Fædrelandsvennen	2,296
Bergens Tidene	1,675
Aftenposten	923
Total	35,194

Trinn i prosessen

- Hente ut anmeldelser
  - Sortere ut feil (manglende terningkast, duplisert tekst, osv).
  - Splitte samleanmeldelser
- Identifisere og normalisere meta-informasjon:
  - Terningkast, url, dato, kategori, osv
- Normalisere mark-up; felles mellomformat i HTML
- Hente ut relevant tekst
- Preprosessering av tekst:
  - Språkgjenkjenning (med *langid.py*): BM / NN.
  - Setningssplitting, tokenisering, lemmatisering, PoS-tagging, dependensparsing (med UDPipe), representert i formatet CoNLL-U.
  - Inkluderer også scripts for å genere versjon i rå tekst.

Anmeldelser per kategori

Category	# Reviews
Screen	13,085
Music	12,410
Literature	3,530
Products	3,120
Games	1,765
Restaurants	534
Stage	530
Sports	118
Misc	102
Total	35,194

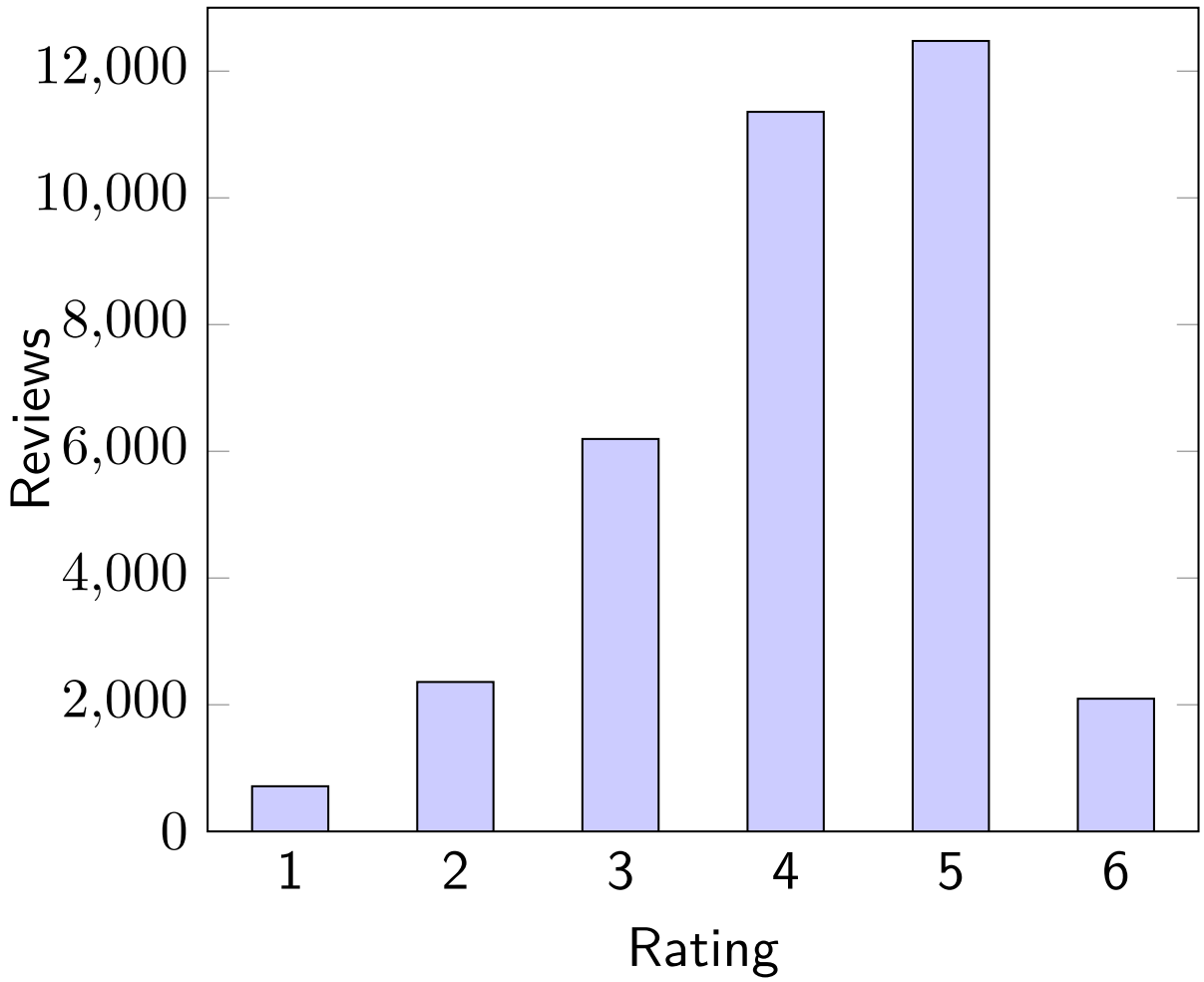
CoNLL-U

1	Men	men	CCONJ	3	cc
2	ikke	ikke	ADV	3	advmod
3	forvent	forvente	VERB	0	root
4	god	god	ADJ	5	amod
5	brukervennlighet	brukervennlighet	NOUN	3	obj
6	,	$,	PUNCT	3	punct
7	det	det	PRON	8	obj
8	får	få	VERB	3	conj
9	du	du	PRON	8	nsubj
10	nemlig	nemlig	ADV	8	advmod
11	ikke	ikke	ADV	8	advmod
12	.	$.	PUNCT	3	punct

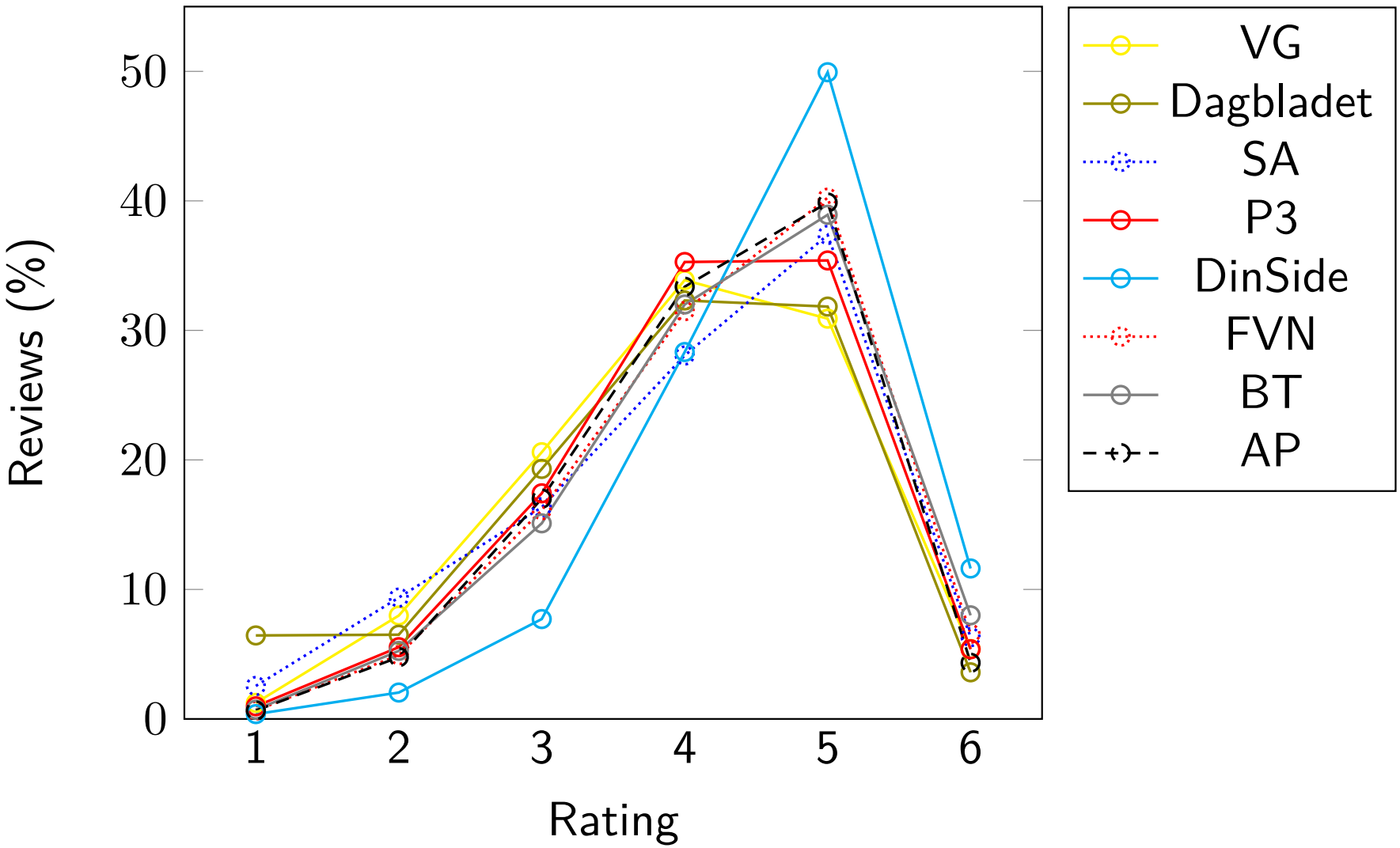
metadata.json


```
"705084": {
    "title": "F-A-N-T-A-S-T-I-S-K",
    "authors": [ "Leif Tore Lindø" ],
    "rating": 6,
    "category": "music",
    "day": 7,
    "month": 12,
    "year": 2015,
    "excerpt": "Det var for sånne konserter Gud fant opp sekseren.",
    "id": 705084,
    "language": "nb",
    "source": "sa",
    "source-category": "konsert",
    "source-id": 203726,
    "source-tags": [ "Anmeldelse", "Musikk", "Konsertanmeldelse" ],
    "split": "train",
    "tags": [ "concert" ],
    "url": "http://www.aftenbladet.no/article/sa-203726b"
  },
```

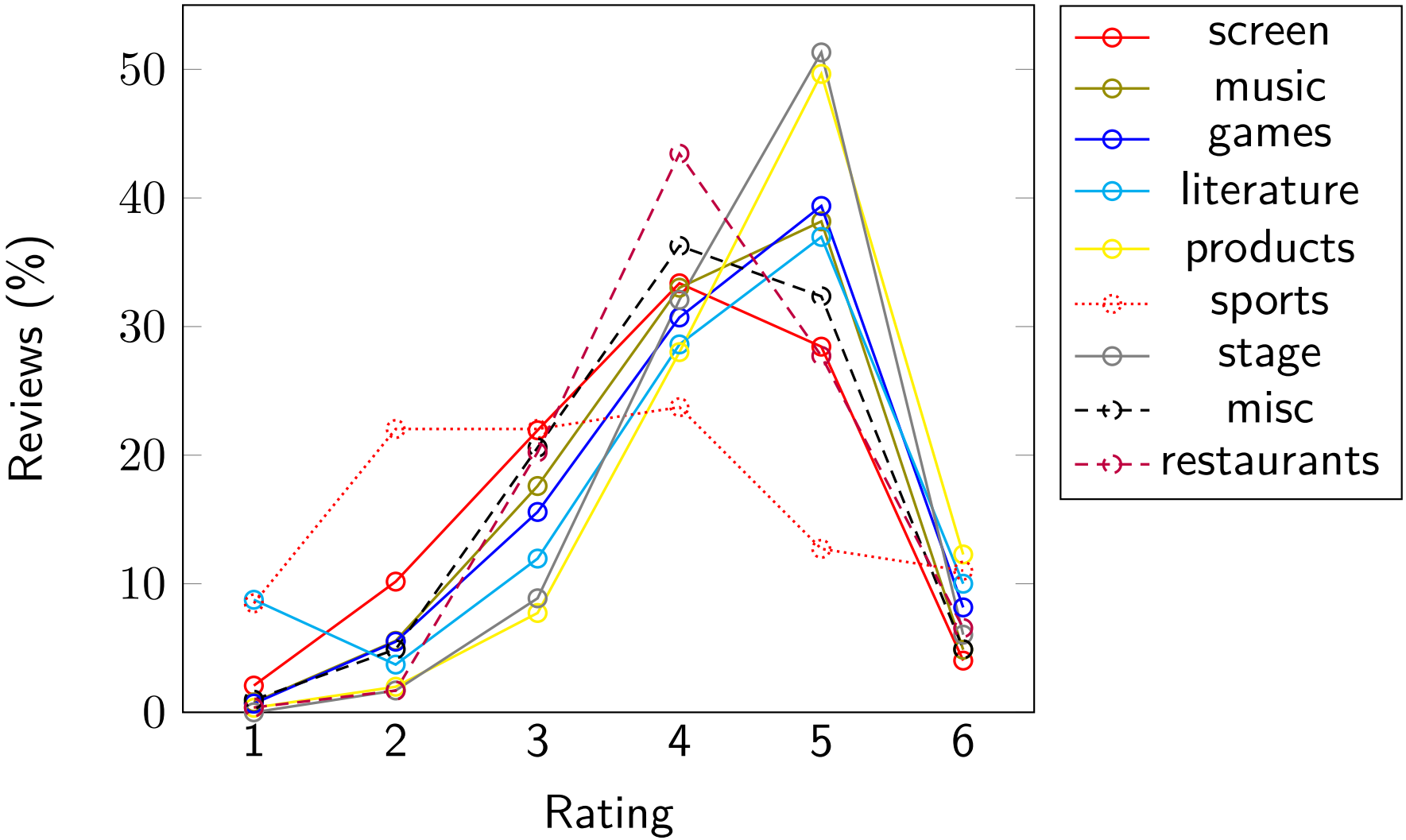
Anmeldelser per terningkast



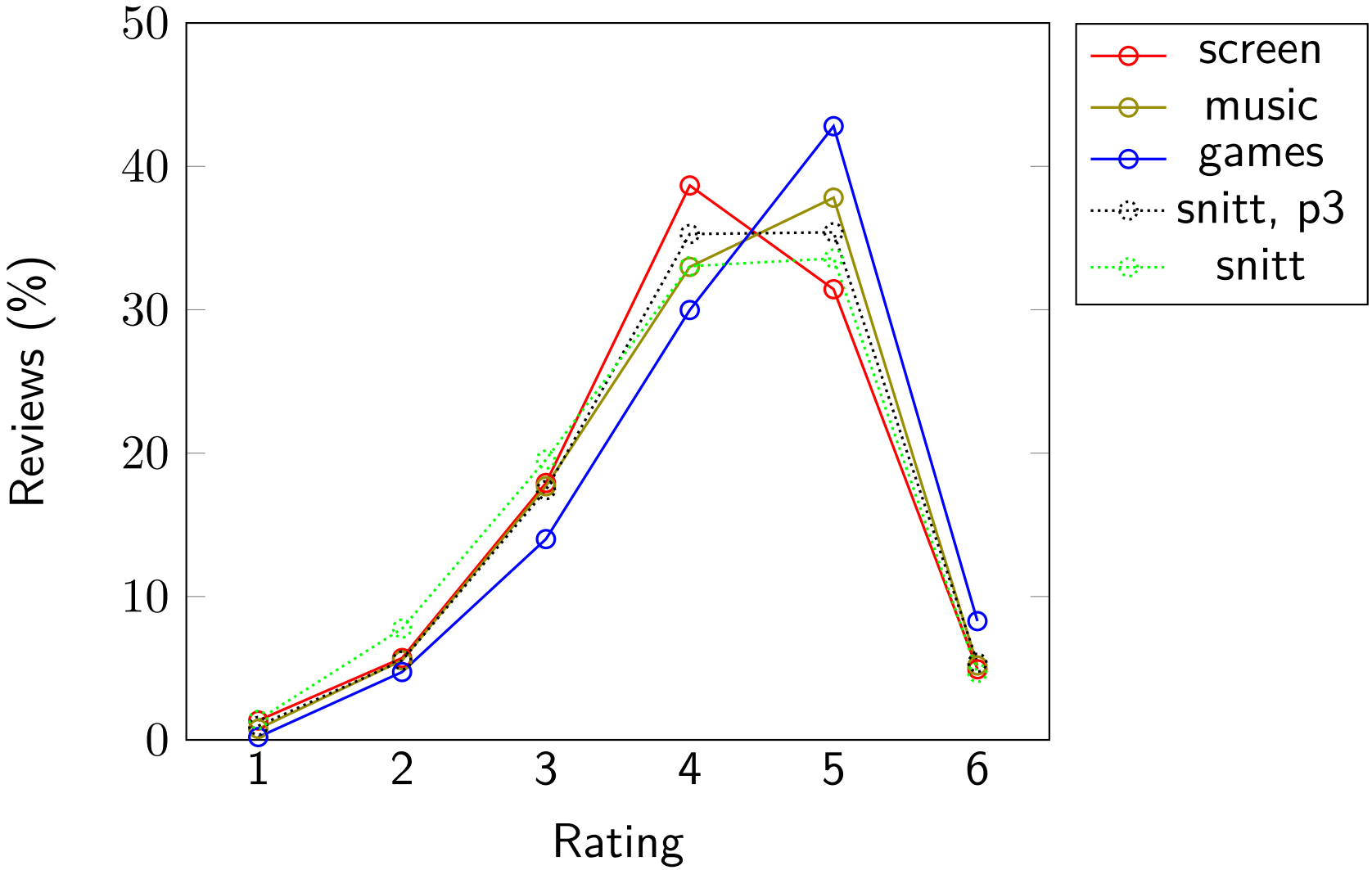
Terningkast per kilde



Terningkast per kategori



Terningkast per kategori: P3 alene



- 5017 P3-anmeldelser over 3 kategorier (screen=2294, music=2216, games=507)
- Snitt er for kun de tre P3-kategoriene men for alle kilder.

Baselines for SA på dokumentnivå

- Inn: dokument / tekst
- Ut: tall 1–6
- Strategier
  - Klassifikasjon
  - Regresjon
- Evaluering
  - Accuracy
  - R^2 score (1 er best, kan være negativ).
- Dokumentrepresentasjon
  - doc2vec (100 dim.)
  - Bag-of-words (BoW)
- Verktøy: Gensim (for doc2vec), Scikit-learn til resten.
- `https://github.com/ltgoslo/norec-baselines`

Foreløpige baseline-resultater

Model	Acc	R^2
BoW+classification	0.50	0.18
BoW+regression	0.46	0.29
doc2vec+classification	0.44	0.07
doc2vec+regression	0.40	0.14
Majority-class	0.37	−0.6
Random choice	0.17	−2.40

- Merk at dette er ment som en bunnlinje for videre eksperimenter.

Fase 2

- Leverte rapport for fase 1 til NFR 10. nov.
- Presentasjon for NFR 4. des.
- Så får vi vite innen jul om vi får finansiering for fase 2
- $=$ 1 postdoc $+$ 1 PhD.
- Plan for fase 2:
  - SA-leksikon
  - Mer finkornet annotasjon
  - Dyp læring for både setnings-nivå (LSTM) og dokument-nivå (CNN).