

Short Text Categorization Exploiting Contextual Enrichment and External Knowledge

Stefano Mizzaro, **Marco Pavan, Ivan Scagnetto, Martino Valenti**

University of Udine, Italy

Disclaimer

- *“Keep it simple, keep it short, and nobody will complain” [Michael Buckland]*
- The Good Presentation Gold Rule



#ShortTxtCateg...

SM, **MP, IS, MV**
uniud, IT

#Outline

- #pbm
- #approach
- #eval
- @home

The problem

- Short texts are growing
- (at least) 2 reasons
 - Twitter 140 limit
 - Mobile devices, input limitations
- Categorization of short texts, or #ShortTxtCateg

#ShortTxtCateg: why it is useful

- To understand what the txt is about
 - #socceroos: easy
 - Goalkeeper did a good job today: difficult (which team? Which “today”?)
 - “I hate that referee”
 - “I hate that referee... He did not understand my paper”
- We focus on Tweets, but not only (facebook status & comments, txt messages, ...)

#ShortTxtCateg: why difficult

- Not enough data
 - Short sentences
 - Abbreviated words, new coined acronyms
- Typos, misppelings, grammar wrong is often
- Time, ephemeral content
- Ambiguity, Disambiguation is more difficult

Damiano Spina @damiano10 · 11h

POI mentions in Twitter are likely to be ambiguous (~50% unigrams) [#microblog](#)
[#SIGIR2014](#)

[Details](#)

Damiano Spina @damiano10 · Jul 6

Jet lag is almost under control. Gold Coast is simply amazing! [#sigir2014](#)

[Details](#)

#ShortTxtCateg: why difficult

- Not enough data
 - Short sentences
 - Abbreviated words, new coined acronyms
- Typos, misspellings, grammar wrong is often
- Time, ephemeral content
- Ambiguity, Disambiguation is more difficult
- #hashtags: potentially useful, but not "normal words"
- Combination: #WFT?!

Combination: #WFT?!

- #WTF = Whom To Follow
- but also...
 - #WTF = What the F*&%
- or, for IR researchers,
 - #WTF = Where is The F^%\$#& data?

Aim

- Find categories/labels that describe the general topic of a short text
- More specifically:
 - Select the **Wikipedia categories** that best describe a **tweet**

1. Literature	Letteratura
2. Economics	Economia
3. History	Storia
4. Philosophy	Filosofia
5. Science	Scienza
6. Entertainment=(Hobby, Entertainment)	Intrattenimento=(Hobby, Intrattenimento)
7. Finance	Finanza
8. Politics=(Politics, Law)	Politica=(Politica, Diritto)
9. Food and drink	Alimentazione
10. Video games	Videogiochi
11. Computer science	Informatica
12. Health and fitness=(Health, Physical fitness)	Salute e fitness=(Salute, Fitness)
13. Fashion	Moda
14. Medicine	Medicina
15. Music	Musica
16. Engines=(Automobiles, Auto racing, Motorcycle sport)	Motori=(Automobili, Automobilismo, Motociclismo)
17. Photo and Video=(Photography, Film)	Foto e Video=(Fotografia, Cinema)
18. Sports	Sport
19. Places=(Tourism, Geography, Travel)	Luoghi=(Turismo, Geografia, Viaggi)
20. Meteorology	Metereologia

Table 1: Wikipedia categories used in our systems, in English and Italian. The notation $X=(Y,Z,\dots)$ denotes the labels we made to group categories about related topics.

Outline

- #pbm
- #approach
- #eval
- @home

Our approach

- Exploiting Wikipedia
 - Search engine
 - Article/category labels
 - Category relationships
- Enrichment
 - Exploiting search engines
- Time aware

Categories selection

- We select the Wikipedia articles by search
- We extract their categories
- We browse the category graph
- We pick the nearest ones

3 versions of a system

1. W2C
2. FEL
3. WEL

3 systems

	Wikipedia pages	Wikipedia SE	Wikipedia category tree	Text Enrichment	Dynamic term selection
1. W2C	Y	Y	Y	N	N
2. FEL	Y	Y	Y	Y	N
3. WEL	Y	Y	Y	Y	Y

1. W2C

- Step 1: Article selection
 - Query definition, by using **bi-grams** from short text
 - Article retrieval process (ranked by Wikipedia search engine)
 - Article re-weighting process, (exploiting their positions in the ranking)
 - Final articles list with distinct entries (by performing all queries and summing the scores)
- Step 2: Label selection
 - Wikipedia categories extraction (for each article)
 - Article-Macro-category relationship definition (based on **shortest paths**)
 - Wikipedia Macro-categories selection (based on our ranking function)
 - Final set of **5 labels**, based on selected Macro-categories

Workflow

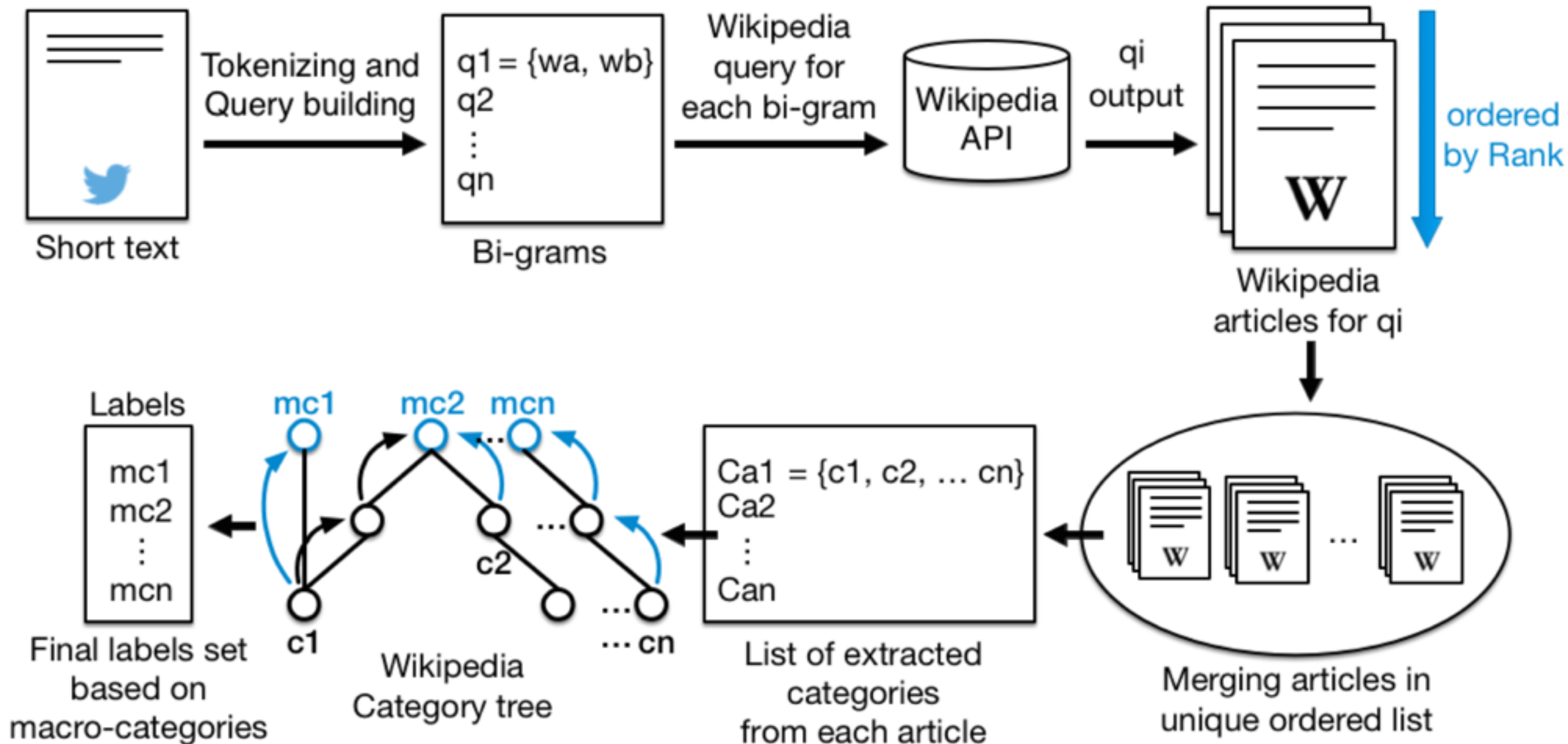


Figure 1: Workflow of the labelling process.

2. FEL

- Enters (short) text enrichment
- The short txt is augmented with some other terms

Workflow

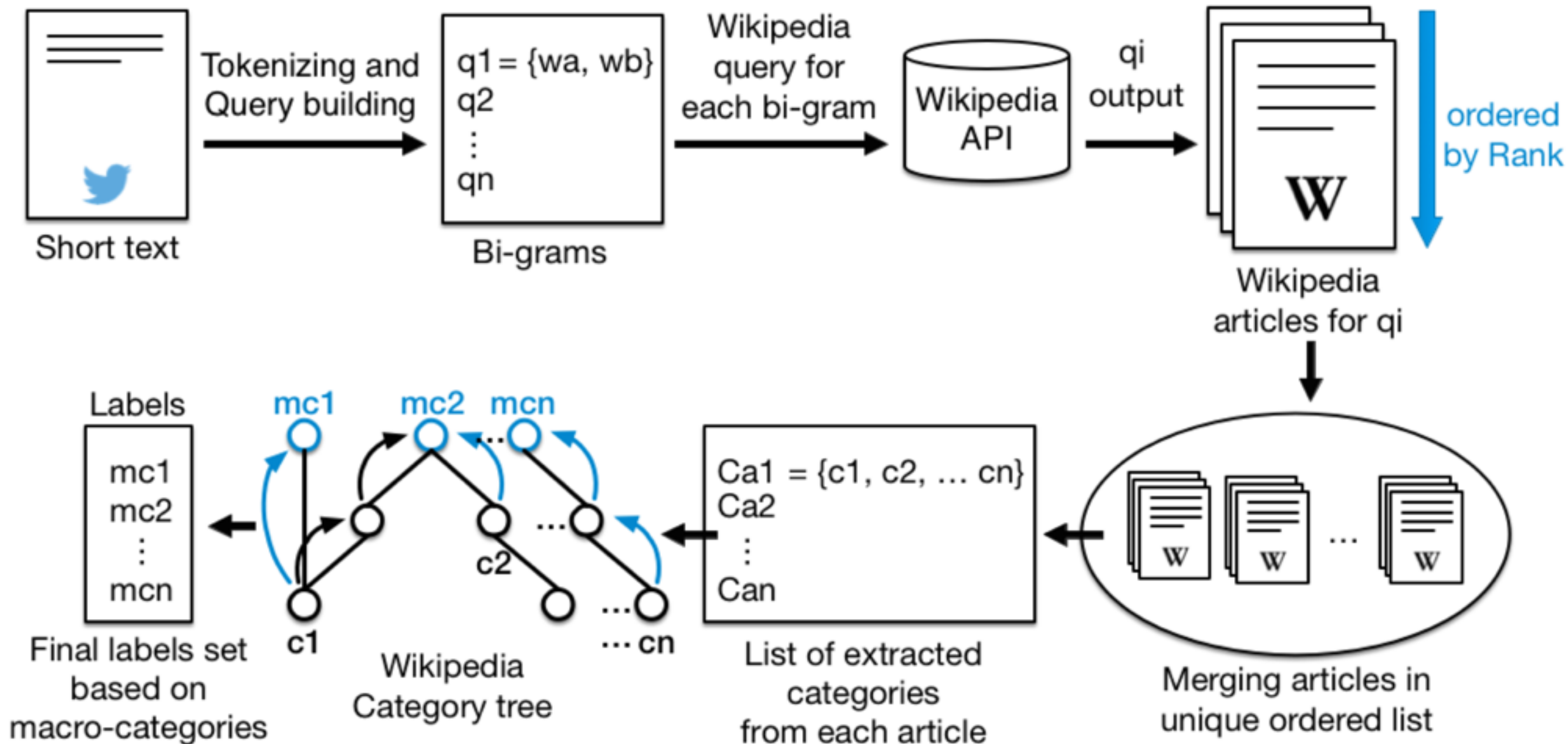


Figure 1: Workflow of the labelling process.

Workflow

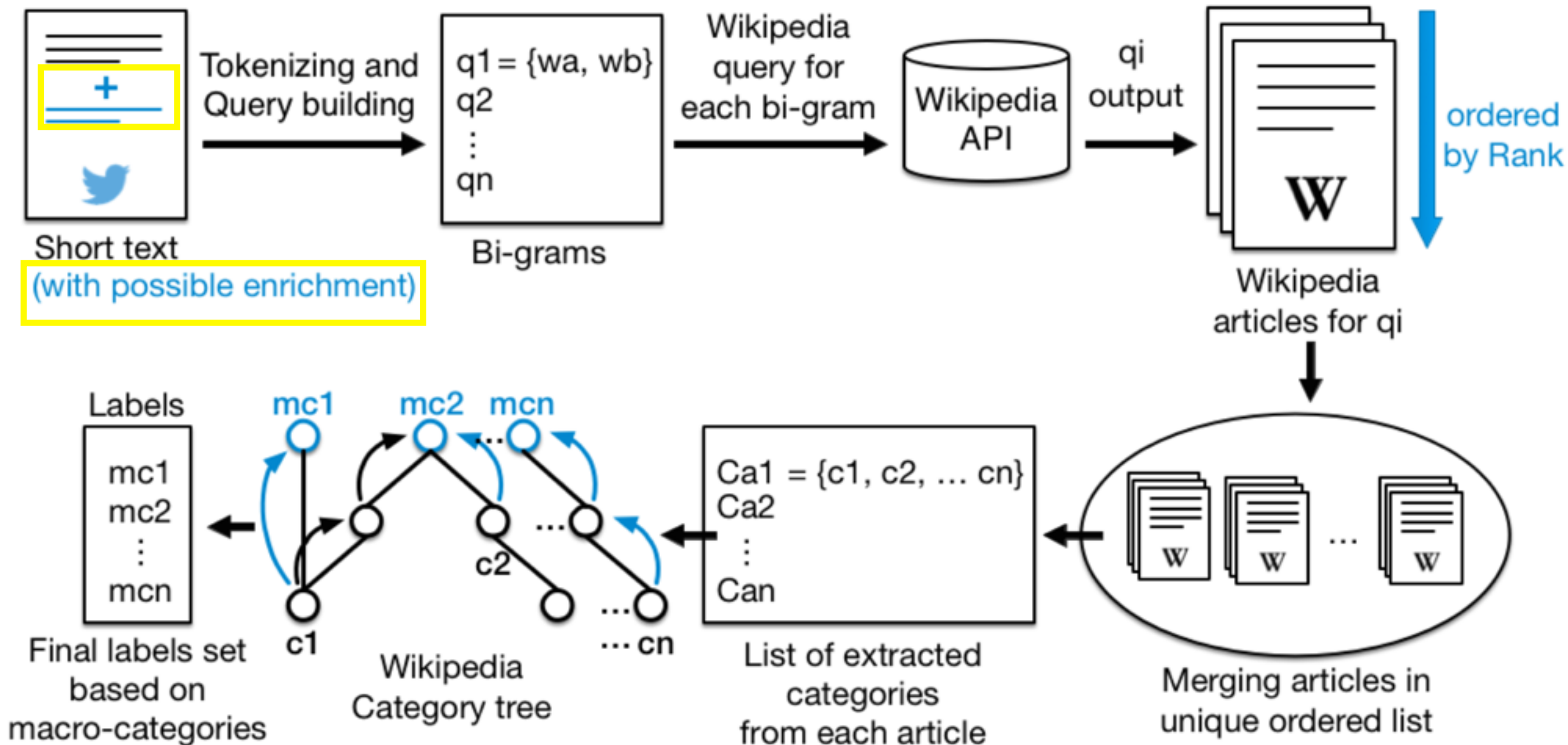


Figure 1: Workflow of the labelling process.

Text enrichment

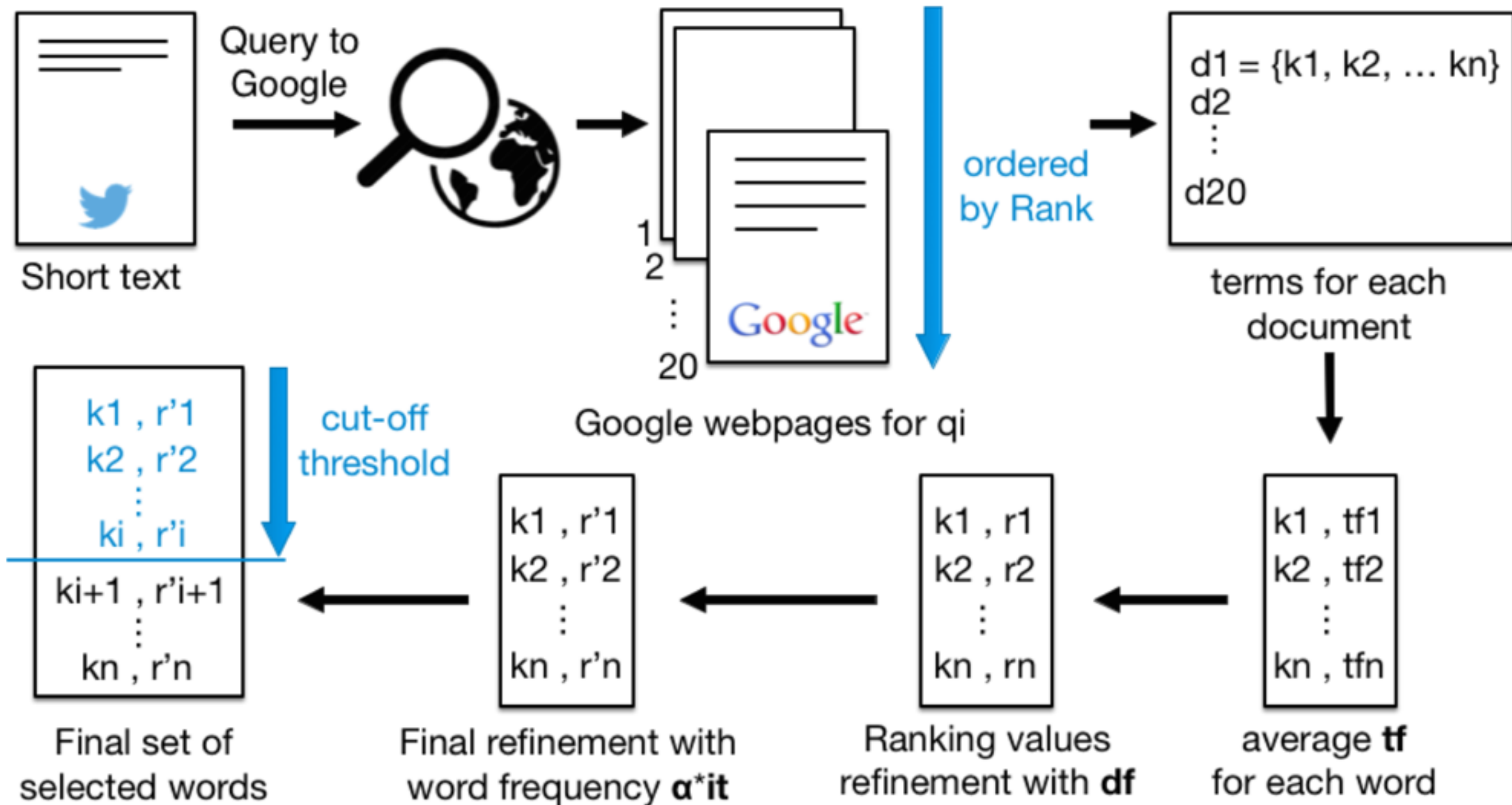


Figure 3: Workflow of the enrichment process.

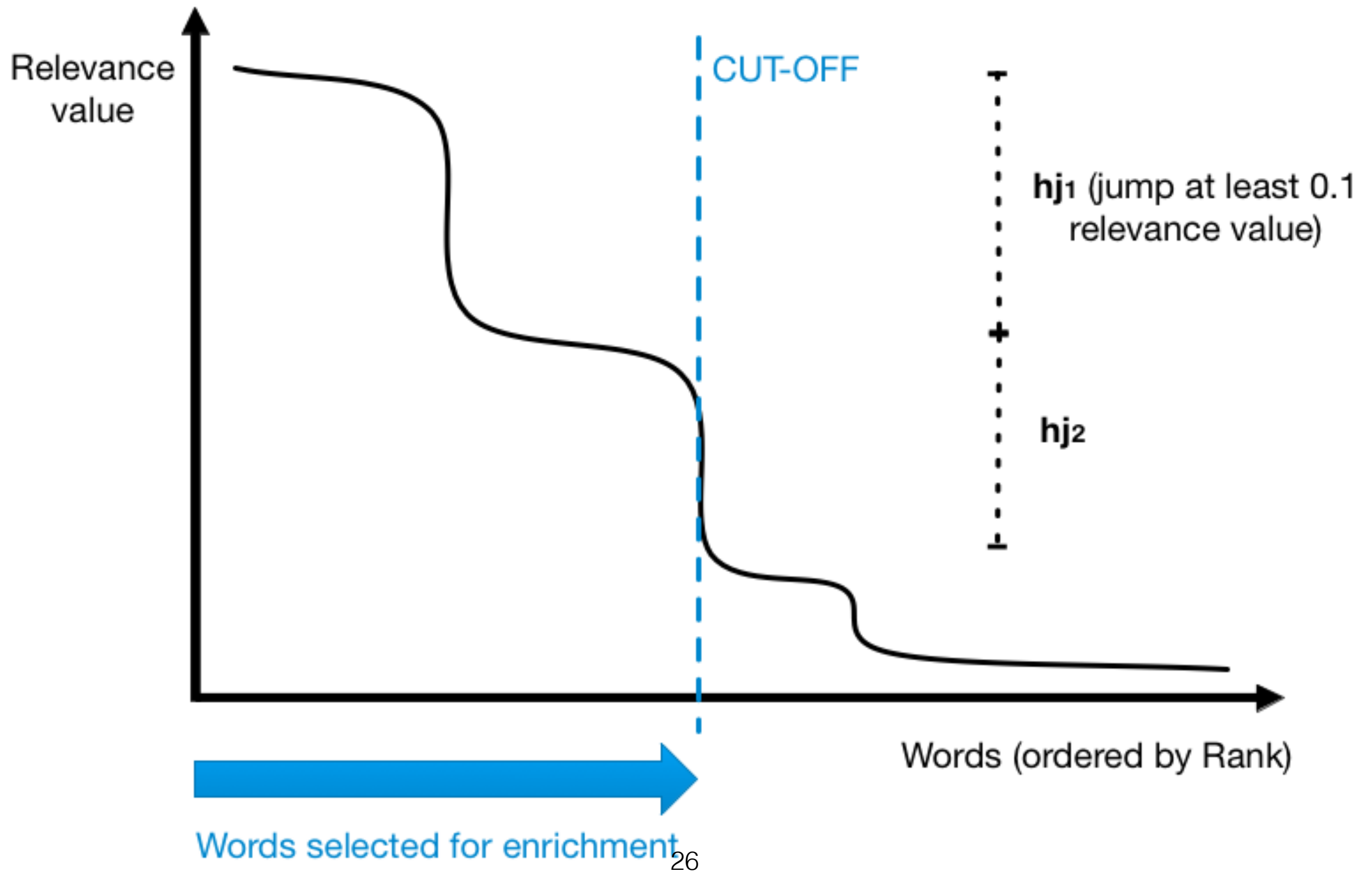
Now, Time

- To be timely is important. I should have said that earlier...

Now, Time

- To be timely is important. I should have said that earlier...
- We query google right after the tweet
- Well actually a few hours (6) after the tweet.

3. WEL



Outline

- #pbm
- #approach
- #eval
- @home

Experimental evaluation

- 3 versions of the system (W2C, FEL, WEL), which is better?
- 20 labels/categories
- 10 twitter accounts
 - 30 tweets
- Assessments by 66 people

Assessing

- Participant was shown a set of labels generated by a system
 - *“Is this set of labels good for describing the topic of the tweet?”*
- 5 levels scale (1=worst, 5=best)
- Usual random shuffling, avoiding learning effects, etc.

Results

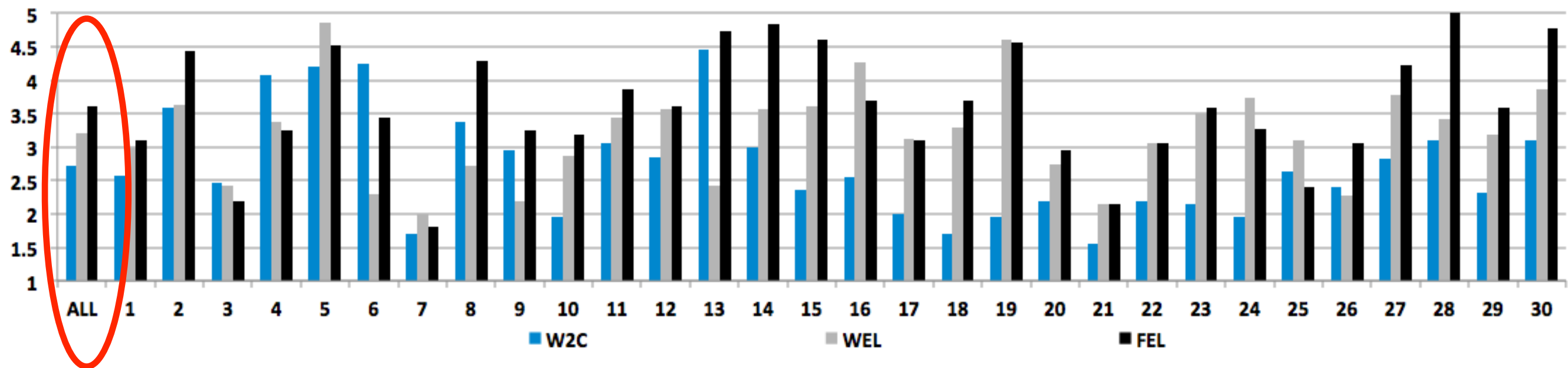


Figure 4: Average rating for each short text

- Statistically significant
- High variance over tweets

Results

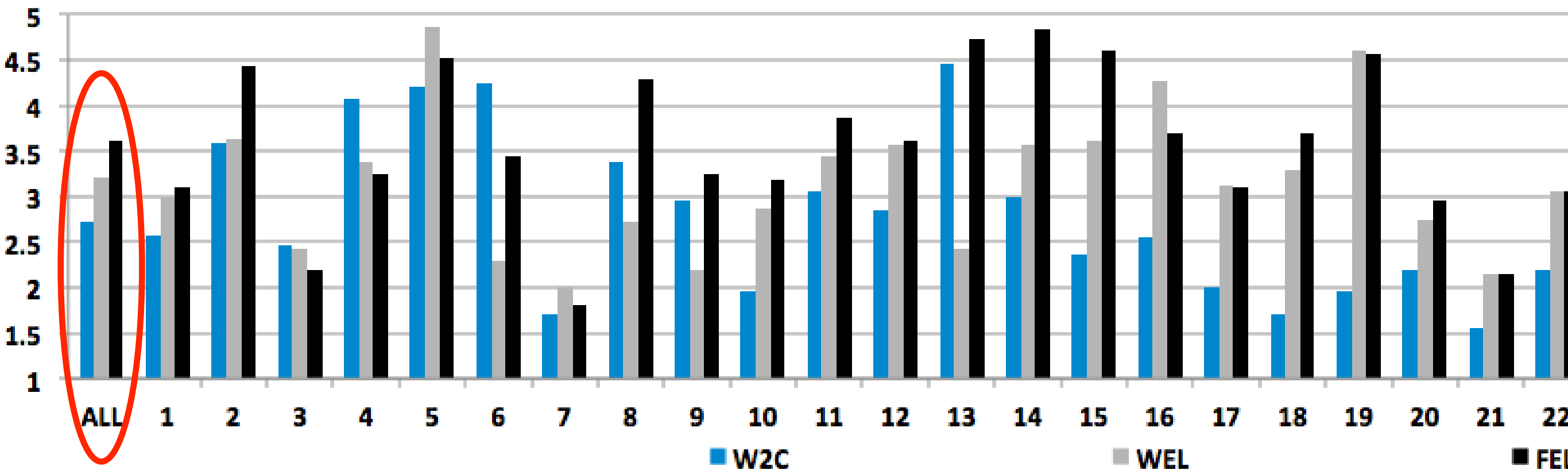
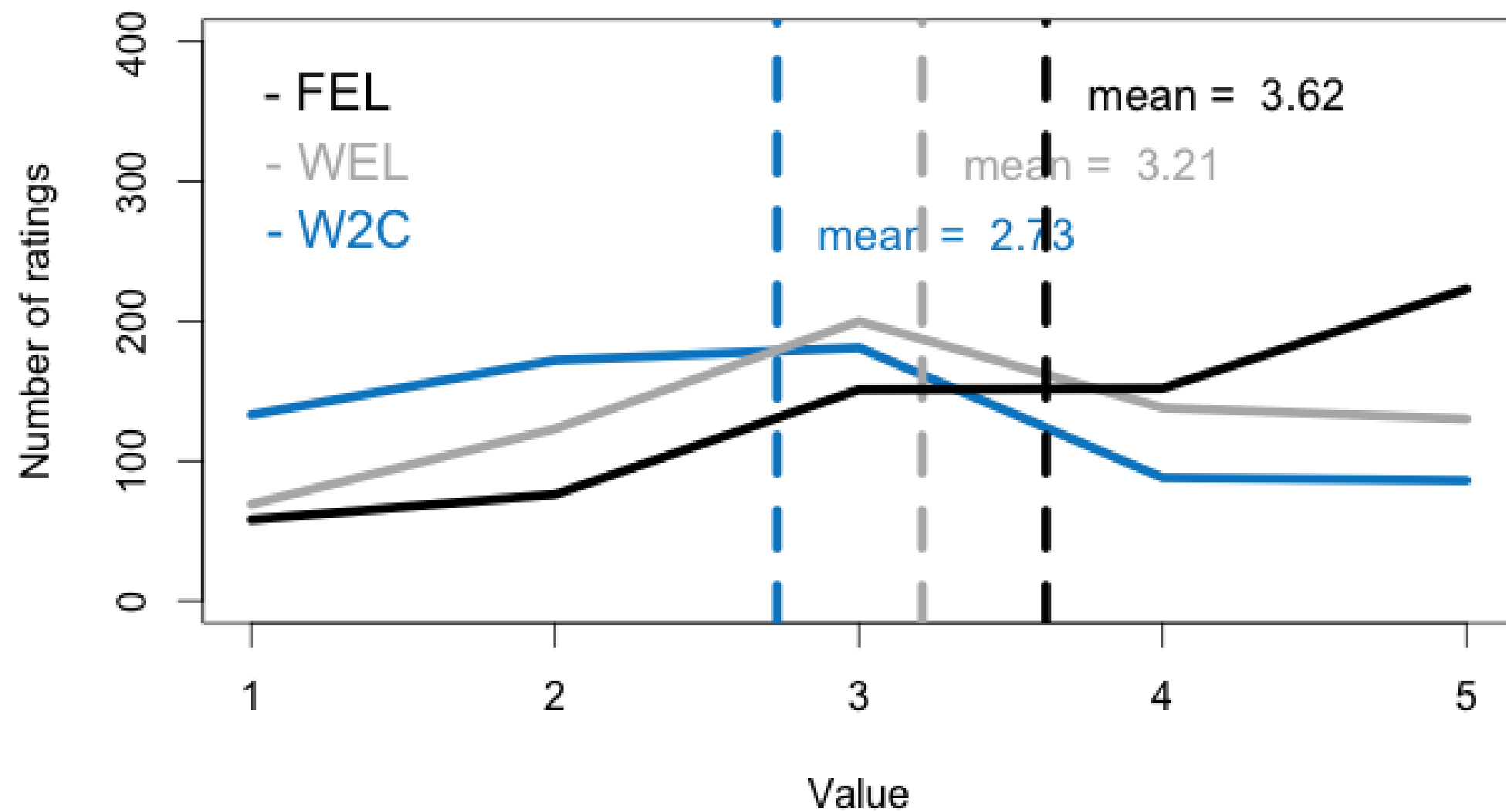


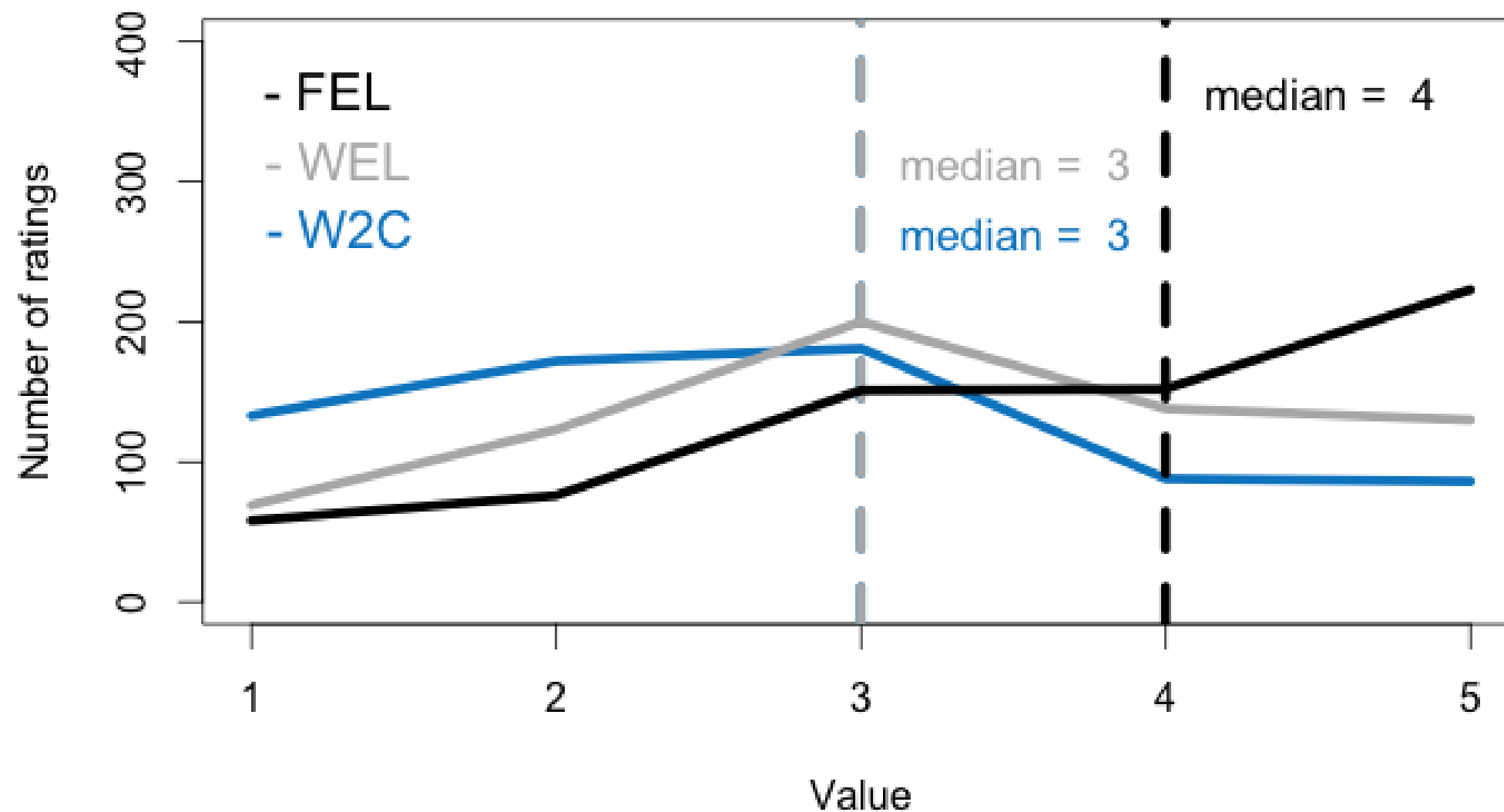
Figure 4: Average rating for each short text

- Statistically significant
- High variance over tweets

Rating distributions



Rating distrib w/ medians



Outline

- #pbm
- #approach
- #eval
- @home

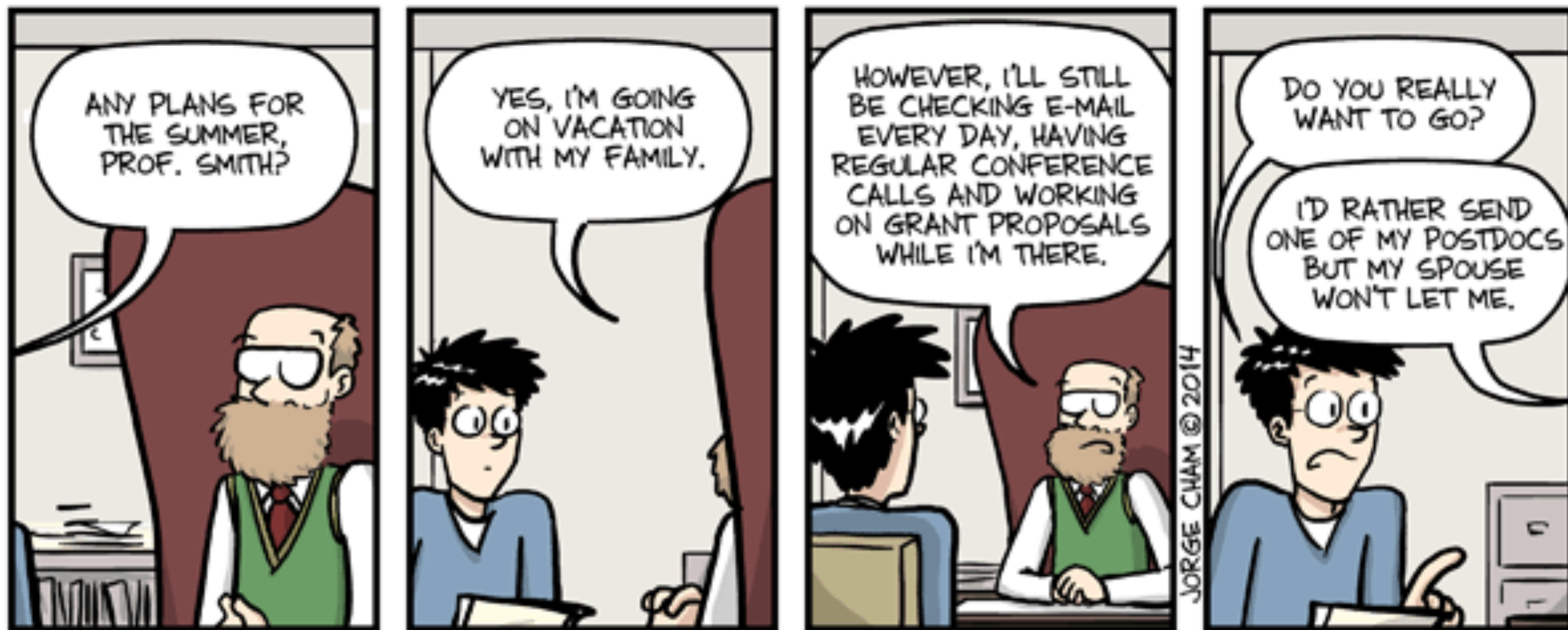
Conclusions

- #ShortTxtCateg
- @timeaware
- w/ or w\ txt enrichment
- txt enrichm seems useful
 - 2. FEL better than 3. WEL

Future work

- #WTF?
- Too much to be listed here
- Plenty of space for improvement

#Tnx!



WWW.PHDCOMICS.COM