Nepal College of Information Technology

Affiliated to Pokhara University

Final Defense

on

Generating Opinion Lexicon for Nepali using Microblogs

Presented by

Sawan Vaidya

(15526)

Supervisor: Asst. Prof. Bal Krishna Bal

Background

Sentiment Analysis (Opinion Mining)

Definition from Google:

noun

the process of computationally identifying and categorizing opinions expressed in a piece of text, especially in order to determine whether the writer's attitude towards a particular topic, product, etc. is positive, negative, or neutral.

"companies have key lessons to learn about harnessing the power of social media and sentiment analysis"

Background

Micro-Blogs

Definition from Google:

noun

a social media site to which a user makes short, frequent posts.
"she announced her retirement on her microblog, bidding farewell to the sport"

Examples: Twitter, Youtube Comments, Ecommerce Reviews, Article Comments etc

Background

Lexicon – 2 Word Lexicon

word	positive	negative
राम्रो	0.9	0.1
नराम्रो	0.1	0.9

=

word	Score
राम्रो	0.8
नराम्रो	-0.8

$$\text{score} = positive - negative$$

$$1 = positive + negative$$

Background

Lexicon – Emoji Sentiment Data[4]

Emoticon	Positive	Negative	Neutral	Emoticon	Positive	Negative	Neutral
	0.08	0.79	0.13		0.68	0.05	0.27
	0.08	0.79	0.13		0.60	0.04	0.36
	0.08	0.79	0.13		0.66	0.11	0.23
	0.06	0.75	0.19		0.62	0.09	0.29
	0.10	0.78	0.12		0.61	0.09	0.30
	0.13	0.77	0.09		0.63	0.11	0.26
	0.08	0.72	0.20		0.62	0.11	0.27
	0.15	0.77	0.08		0.63	0.13	0.24
	0.13	0.73	0.15		0.62	0.12	0.25
	0.12	0.72	0.16		0.59	0.12	0.29

Background

Lexicon – 32 Word Manual

word	positive	negative	neutral	word	positive	negative	neutral	word	positive	negative	neutral
राम्रो	0.9	0.05	0.05	शान्ति	0.9	0.05	0.05	जबरजस्ती	0.05	0.9	0.05
स्वच्छ	0.9	0.05	0.05	शुद्ध	0.9	0.05	0.05	हस्तक्षेप	0.05	0.9	0.05
चम्किलो	0.9	0.05	0.05	स्वस्थ	0.9	0.05	0.05	नकारात्मक	0.05	0.9	0.05
सभ्य	0.9	0.05	0.05	नियमित	0.9	0.05	0.05	फटाहा	0.05	0.9	0.05
सफा	0.9	0.05	0.05	व्यवस्थित	0.9	0.05	0.05	लुटेरा	0.05	0.9	0.05
सक्षम	0.9	0.05	0.05	नराम्रो	0.05	0.9	0.05	चोर	0.05	0.9	0.05
शान्त	0.9	0.05	0.05	भ्रष्ट	0.05	0.9	0.05	वाक्क	0.05	0.9	0.05
पौष्टिक	0.9	0.05	0.05	सुस्त	0.05	0.9	0.05	आतंक	0.05	0.9	0.05
ताजा	0.9	0.05	0.05	हचुवा	0.05	0.9	0.05	तोडफोड	0.05	0.9	0.05
निष्पक्ष	0.9	0.05	0.05	अनियमित	0.05	0.9	0.05	अपहरण	0.05	0.9	0.05
विश्वसनीय	0.9	0.05	0.05	झुर	0.05	0.9	0.05				

Problem Statement

STATEMENT

"A basic ingredient of **Sentiment Analysis** that uses Bag-of-words approach is an **Opinion Lexicon** containing words/phrases along with their positivity and negativity ratings. Many methods have been suggested for building such a lexicon. In this thesis a method **utilizing micro-blogs** (user comments from Facebook pages) is suggested. While, this lexicon will be **domain specific**, it will be a **first in the Nepalese context** and can lay the foundation for development of Sentiment Analysis resources. Major **challenges** are that data generated by users on the Internet is **subject to informal language**. Also **misspellings**, **abbreviations**, **emoticons** and **multi-language** content pose additional hurdles. The benefits are that **micro-blogs are richer in subjective content** compared to text found in formally written sources. The thesis will investigate the use existing lexicons (such as emoticon lexicon) and manually generated lexicons as seed lexicon for enriching and expanding them.." – Text from Report

Objectives

- Develop a method of generating opinion lexicon using text available in microblogs
- Refine the strategy for filtering out mistakes and useless terms from the microblog lexicon
- Generate an opinion lexicon for Nepali, specify its pros and cons.

Literature Review

❖ TWO PREVALENT METHODS

- Dictionary Based
  - Uses word relations in dictionaries (like WordNet [3])
  - Eg. SentiWordNet [1]
- Corpus Based
  - Analyses occurrences of words in documents
  - Mostly uses Pointwise Mutual Information between words. Eg: [5]

Literature Review

❖ Primary Reference

Paper: Detecting Sentiment in Nepali Texts [2]

- Translation of SentiWordNet words to Nepali (Bhavanakos)
- Sentiment Clue List
- Nepali Sentiment Corpus

❖ Limitations

- Translated from English Lexicon
- Not representative of language used in Microblogs
- Absence of Polarity Strengths

Literature Review

Other Inspirational Papers

Paper: Sentiment of Emojis [4]

- Polarity strengths for various Emojis

Paper: Sentiment lexicon generation for an under-resourced language [5]

- Selects top patterns in text
- Uses Pointwise Mutual Information between words

Methodology

Example (Text Classification)

Word	Pos (p)	Neg (n)
राम्रो	0.9	0.1
नराम्रो	0.1	0.9

Primary Lexicon

Use

Classification

Text	Average score(p)	verdict
म राम्रो र सफा	0.8	Positive
म नराम्रो र फोहर	-0.8	Negative
म सफा र शुद्ध	0	Unknown
म फोहर र गनाउने	0	Unknown
म राम्रो र नराम्रो	0	Unknown

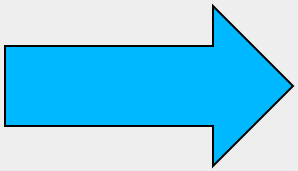
Sentences

Methodology

Example (Counting)

Text	verdict
म राम्रो र सफा	Pos
म नराम्रो र फोहर	Neg

Classification



Count Words

Counts

	Counts	
Word	**Pos**	**Neg**
राम्रो	1	0
सफा	1	0
म	1	1
र	1	1
नराम्रो	0	1
फोहर	0	1

Methodology

Calculate Strength

Word (w)	Pos	Neg	P (w|Pos)	P (w|Neg)	Strength (pos)	Strength (neg)	Score
राम्रो	1	0	0.25	0	1.00	0.00	1
सफा	1	0	0.25	0	1.00	0.00	1
म	1	1	0.25	0.25	0.50	0.50	0
र	1	1	0.25	0.25	0.50	0.50	0
नराम्रो	0	1	0	0.25	0.00	1.00	-1
फोहर	0	1	0	0.25	0.00	1.00	-1
Total	4	4					

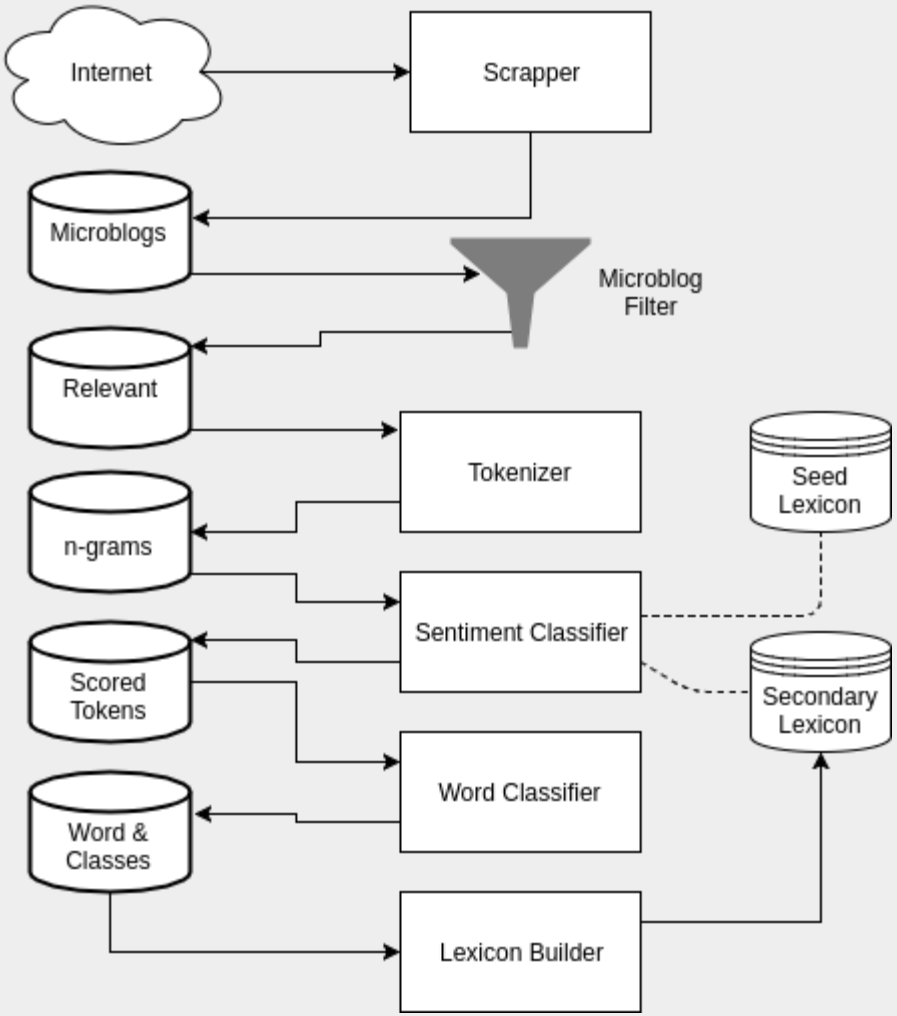
New Words found

Methodology

ASSUMPTIONS

- Bag of Words
- Neutral words occur in both positive and negative contexts
- Positive words occur more with other positive words
- Negative words occur more with other negative words

Methodology



Methodology

Data Collection (facebook pages)

PageID	Comments	PageID	Comments
MaoistCenterNp	2901	bbcsajhasawal	34320
BibeksheelSajha	7590	NepalPolicePHQ	38768
yourMadanKrishnaShrestha	8433	dc.nepal	42433
nepalicongresshq	9864	DrBaburamBhattarai	51384
communistpartynepal	11986	rekhathapa.net	55116
hari.bamsha.acharya	14286	PriyankaKarkiofficial	60660
setopati	15162	BBCnewsNepali	64062
nepal8thwonder	23158	nagariknews	71181
onlinekhabarnews	30267	eKantipur	73286

Methodology

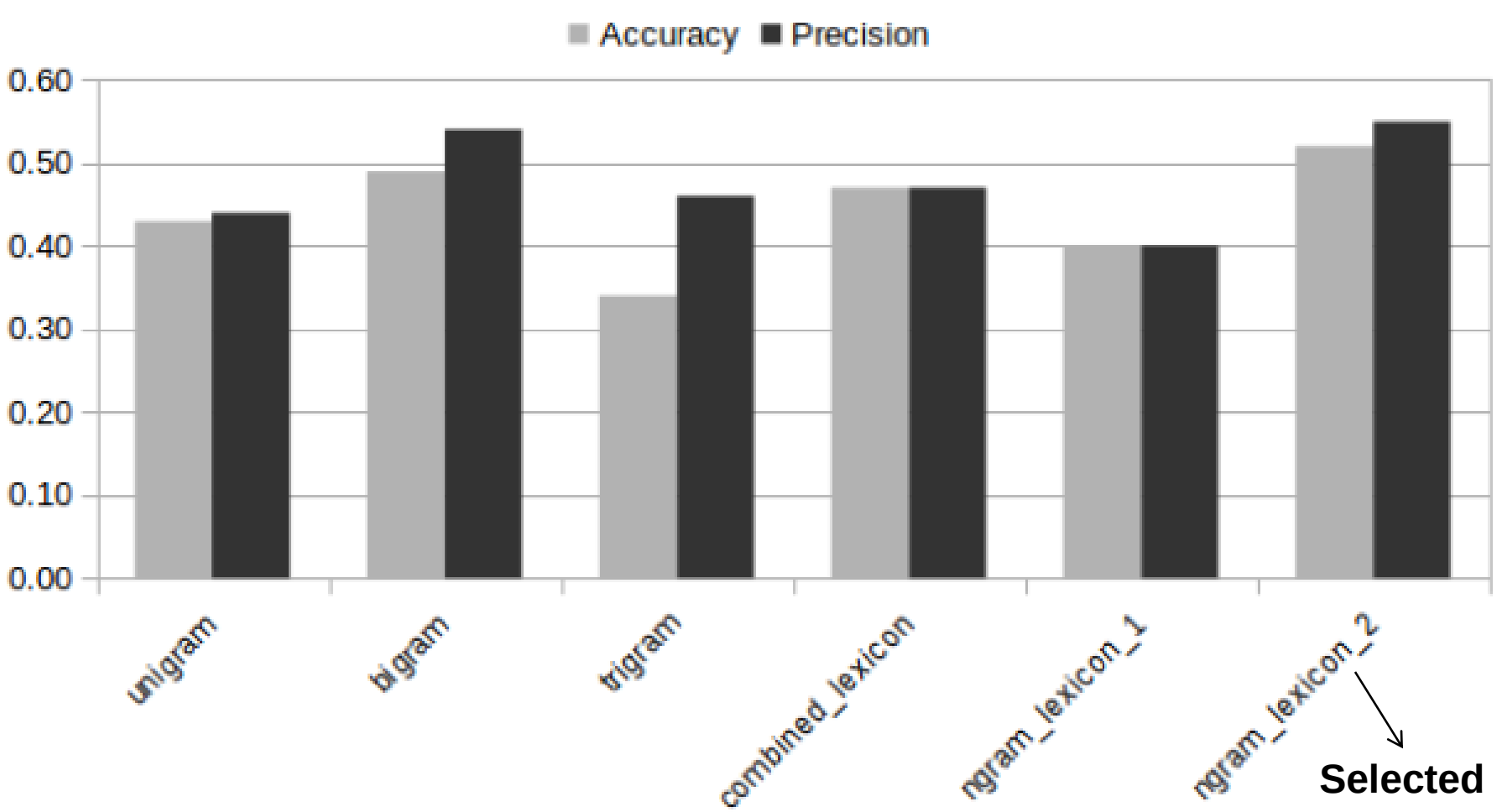
Character Selection

Latin	Devanagari	Emoticons	Count
No	No	No	16694
No	No	Yes	890
No	Yes	No	106506
No	Yes	Yes	3180
Yes	No	No	448296
Yes	No	Yes	11317
Yes	Yes	No	26227
Yes	Yes	Yes	1747
No	**Yes**	**Don't Care**	**109686**

Dataset that was used in bold
1059 Test data was subtracted from it

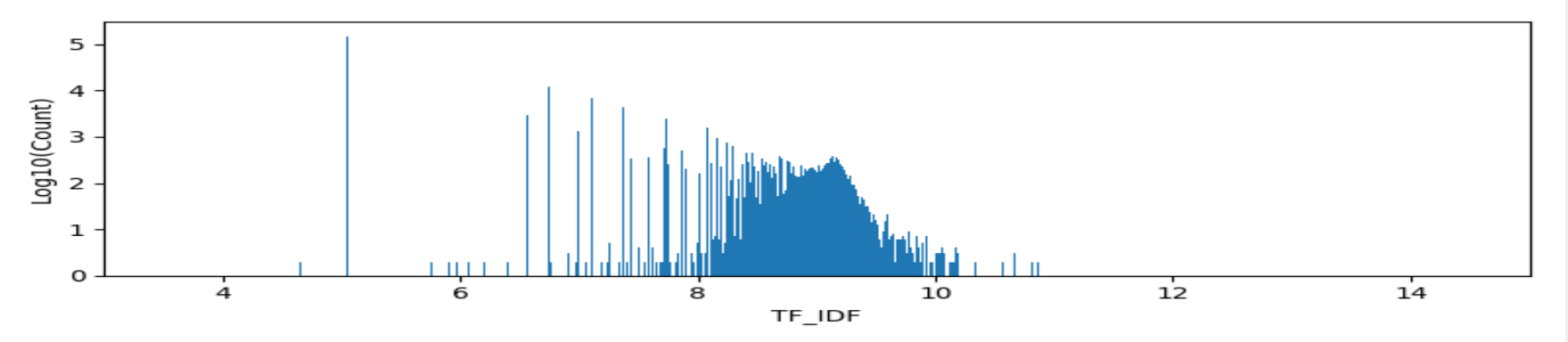
Methodology

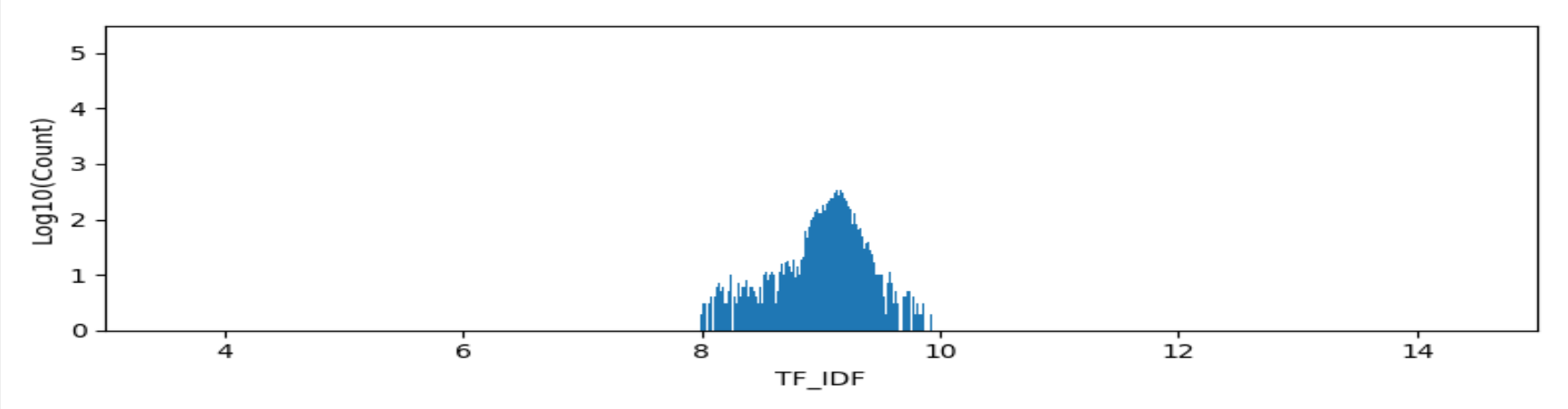
TOKENIZATION METHODS



Methodology

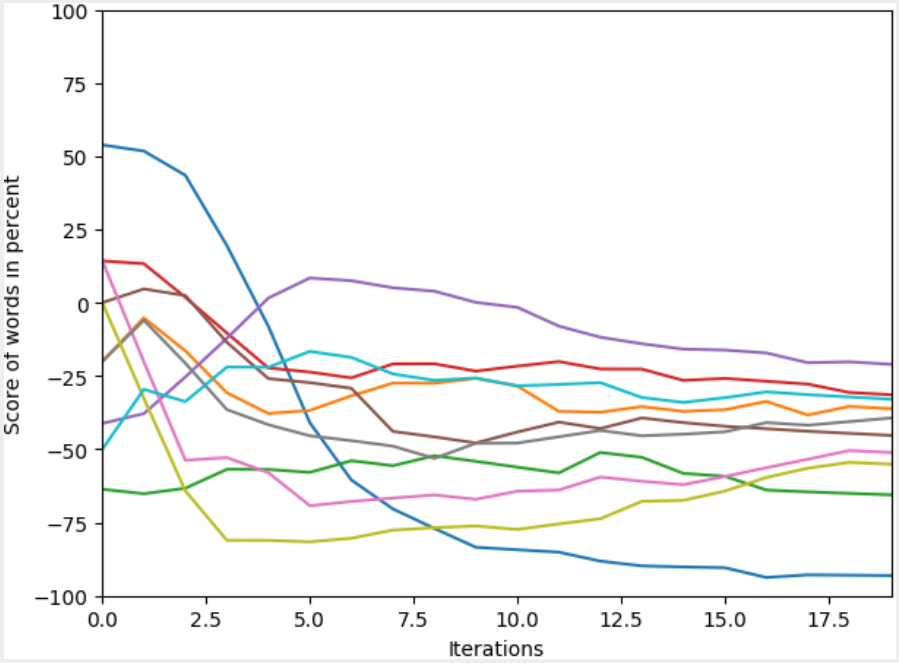
- PRE FILTER : 8 < **TF-IDF** < 10 AND **DOC-COUNT** > 32

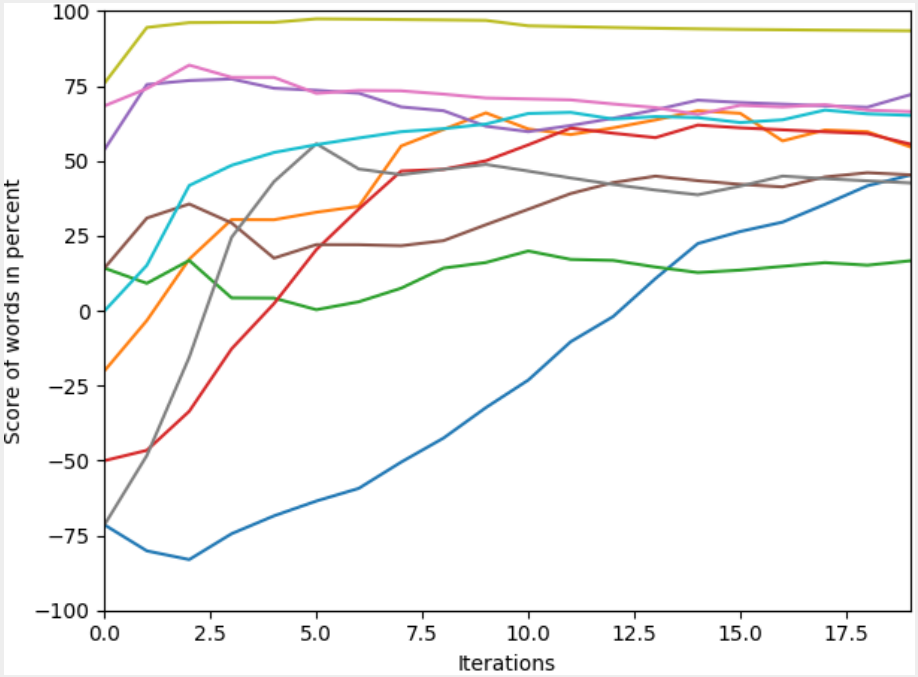




Methodology

Iteration Effect





Methodology

Build Filters : Filters at Lexicon Builder stage

Skewness	$Skewness = \frac{3(mean - median)}{stdev}$
Range	$highest\ score - lowest\ score$
Overall Score	The final score obtained by an ngram
Count	The number of observations for forming a score

Experiments

Experiment	Seed Lexicon
Experiment 1	Emoticon
Experiment 2	32 word
Experiment 3	2 word

Experiments

Initial Parameters

Parameter	Value	Parameter	Value
Devanagari	Mandatory	TF_IDF min	8
Latin	Disallow	TF_IDF max	10
Test Data	Disallow	Doc count min	32
Emoticons	Optional	Iterations	20
TestData	1059	Ngrams	Ngram Method 2
Microblog Count	108566	Build Parameters	See Table 4.2

Range	range>=0.05
Count	count>=Mean/4
Score	|score|>=0.5
Skewness	|skewness|>=Mean/8

Evaluation

Test Data : Sample

Text	Sentiment
म मंगल ग्रह बाट सुन्दै छु	0
प्रधानमंत्री को भारत यात्रा ले नेपाळ को फाइदा होइन नोकसान मात्र हुने छ किनकी देउवा भारत को शिष्य हो। कुनै पनि संधि समझेउता नग्रोस भारत संग	-1
हाम्रो देशको बाढी पहिरो कस्ले हेरिदिने कस्ले बुझिदिने ??	-1
धन्य हो हाम्रो नेपाली सेना सबै थाउ मा उहाँ हरू सहयोग ले नै देस बचि राछ मेरो सलाम छ	1

Evaluation

Test Data: Composition

Total Unique Tokens	10277
Total 1 grams matching TF-IDF criteria	4546
Total 2 grams matching TF-IDF criteria	2195
Total 3 grams matching TF-IDF criteria	246
Positive Comments	353
Negative Comments	353
Unknown and Neutral Comments	353

Evaluation

Confusion Matrix : Random Classification

Class	Precision	Recall	F-score
A	0.34	0.33	0.33
B	0.36	0.35	0.35
C	0.3	0.32	0.31
Overall Accuracy	0.33	Mean Precision	0.35

Evaluation

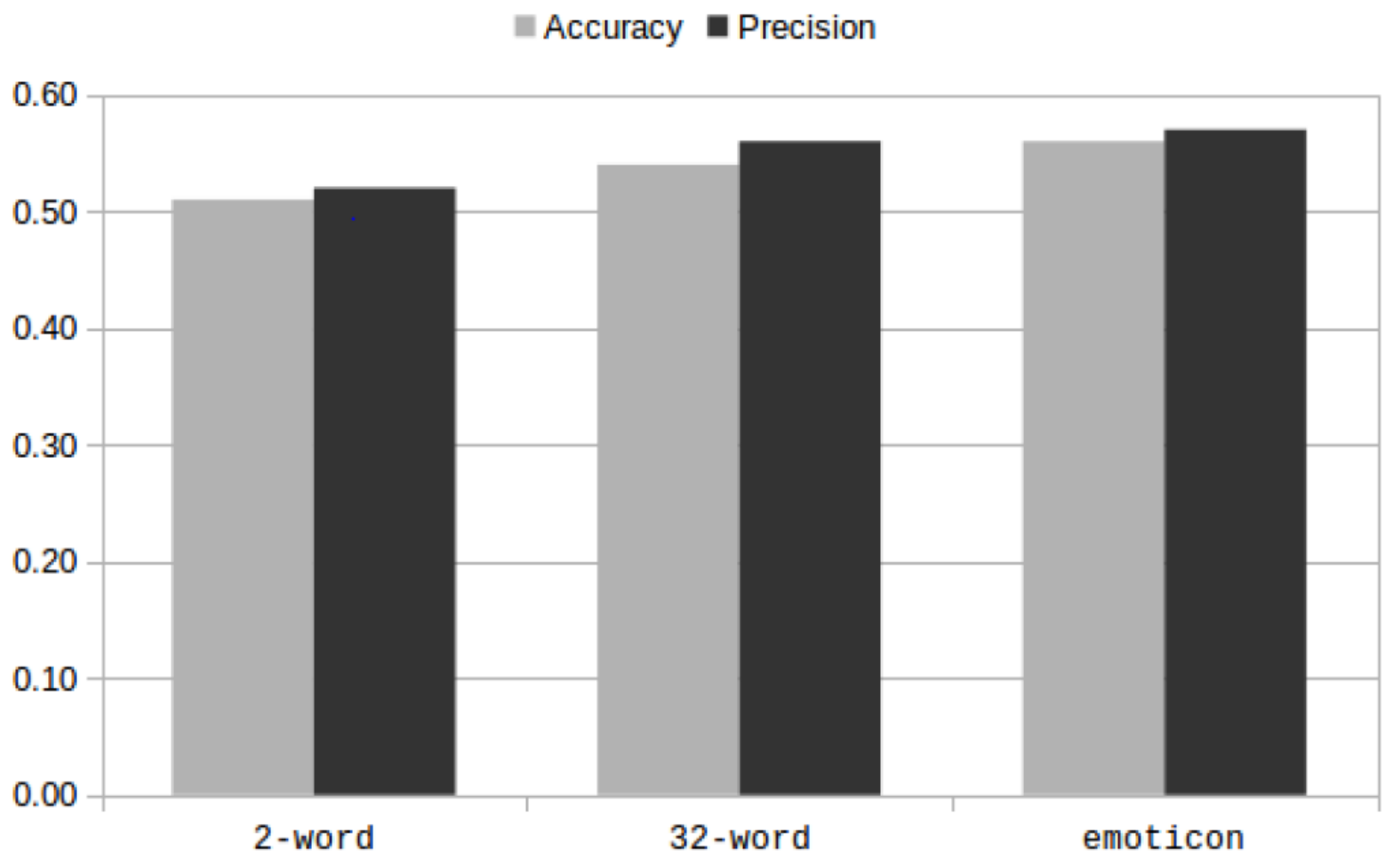
Accuracy and Precision

$$Overall\ Accuracy = \frac{Correct\ Classifications}{Total\ Classifications}$$

$$Mean\ Precision = \sqrt{Precision\ (Pos) * Precision(Neg)}$$

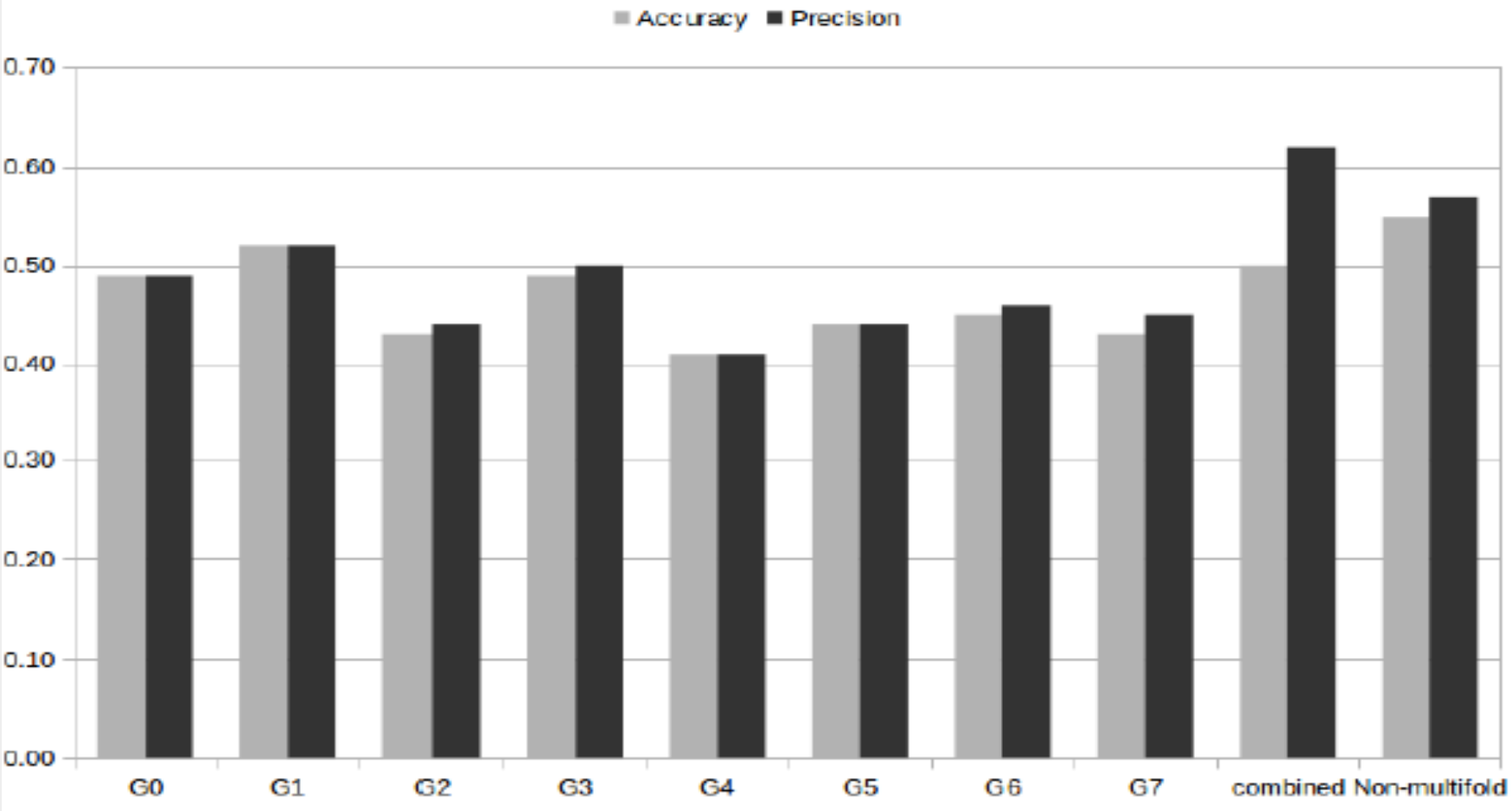
Results

Experiments Compared



Results

Multifold Analysis (Emoticon Seed)



Trigrams (2-word Seed)

Trigrams (32-word Seed)

Trigrams (Emoticon Seed)

Conclusion

PROS

- Suitable for micro-blog domain
- Small number of pre-requisites
- Captures colloquial language features

CONS

- Micro-blog domain dependent
- Unable to distinguish POS
- Difficult to test

Conclusion

Achievements

- Micro-blog domain specific lexicon
- Strategies for refining were studied
- Around 4000 terms could be produced from a simple seed lexicon

Further Work

- Increasing data size can improve accuracy and precision
- Build a domain independent lexicon using WordNet of digital dictionary

References

1 Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In LREC, volume 10, pages 2200--2204, 2010. 4

2 Chandan Prasad Gupta and Bal Krishna Bal. Detecting sentiment in nepali texts: A bootstrap approach for sentiment analysis of texts in the nepali language. In Cognitive Computing and Information Processing (CCIP), 2015 International Conference on, pages 1--4. IEEE, 2015. 6, 26

3 George A Miller. Wordnet: a lexical database for english. Communications of the ACM, 38(11):39--41, 1995. 4

4 Petra Kralj Novak, Jasmina Smailović, Borut Sluban, and Igor Mozetič. Sentiment of emojis. PloS one, 10(12):e0144296, 2015. 7

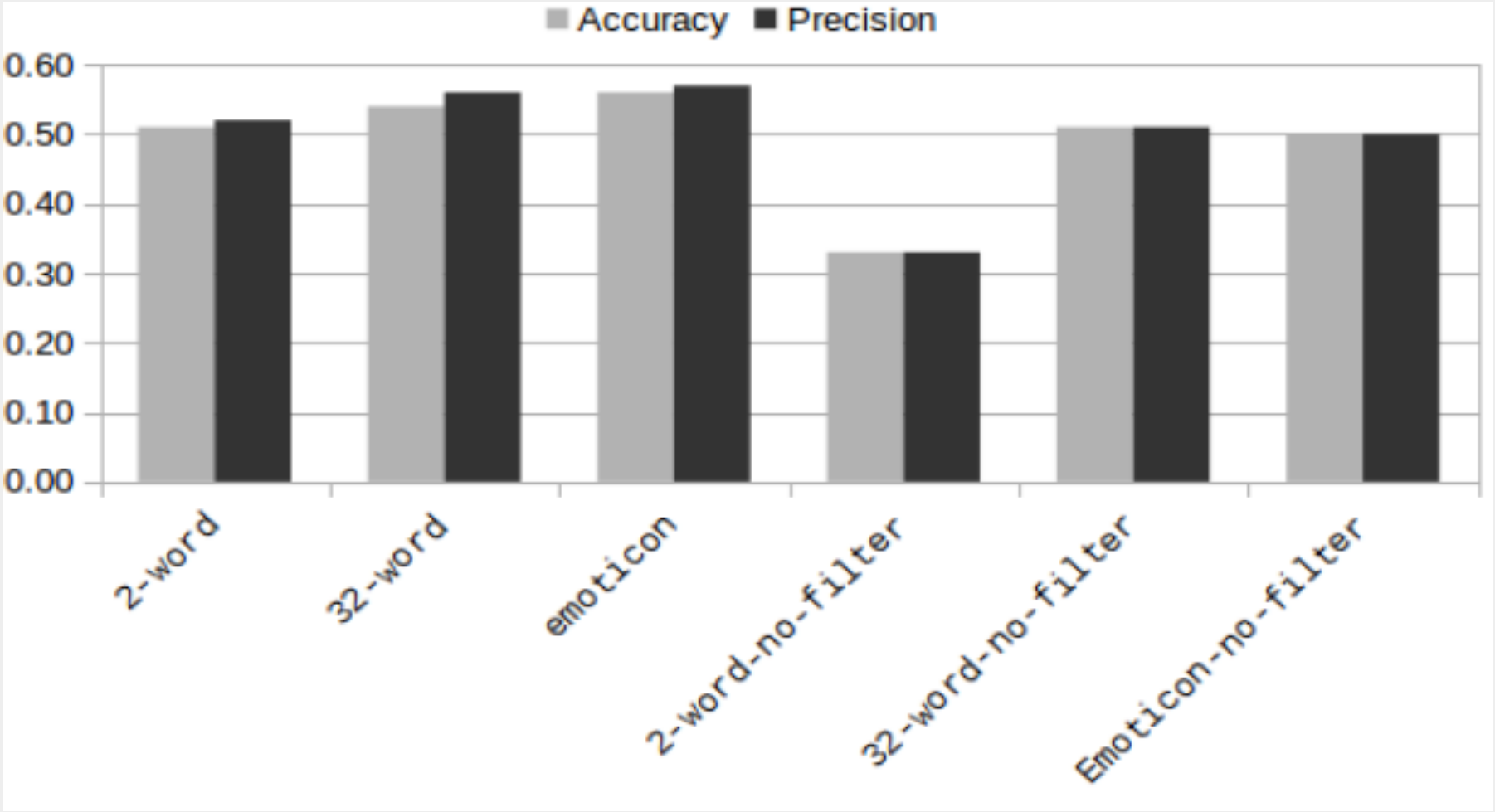
5 Clara Vania, Moh. Ibrahim, and Mirna Adriani. Sentiment lexicon generation for an under-resourced language. Int. J. Comput. Linguistics Appl., 5(1):59--72, 2014. 6

Questions

why
when
who
where
what
why
QUESTIONS
who
how
where
who
what
how
why
when

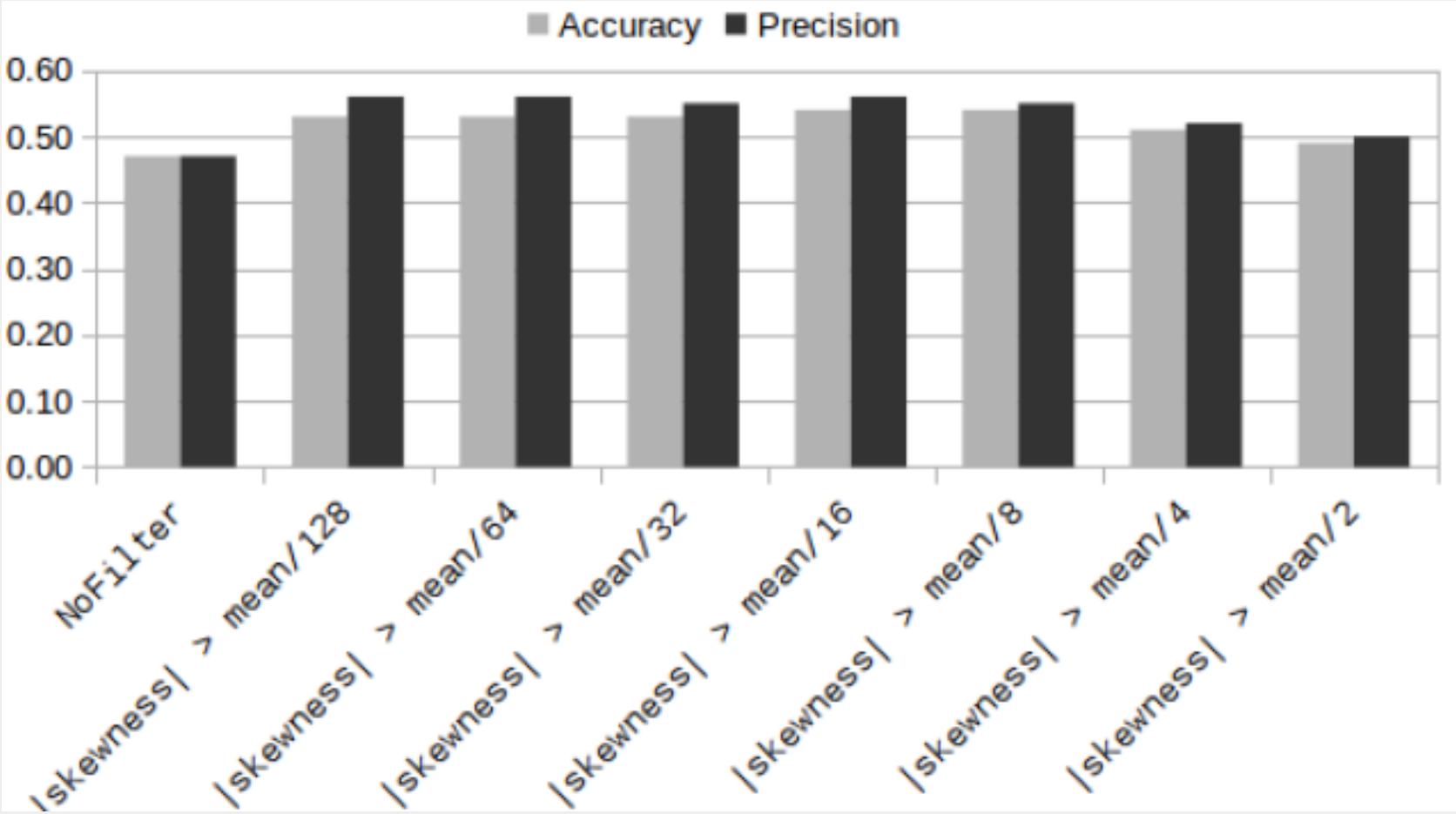
Build Filters

Effect of Filters



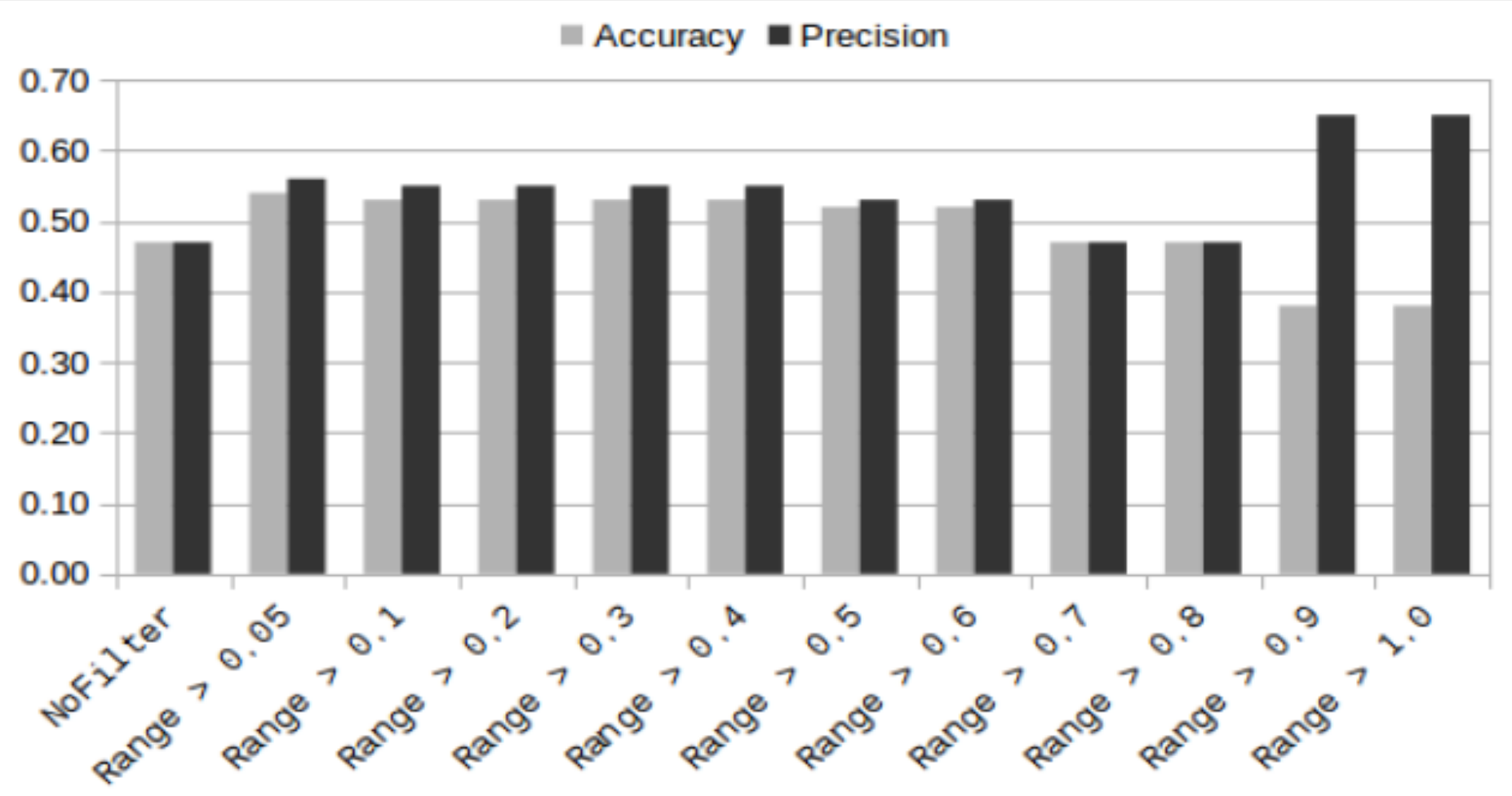
Build Filters

SKEWNESS



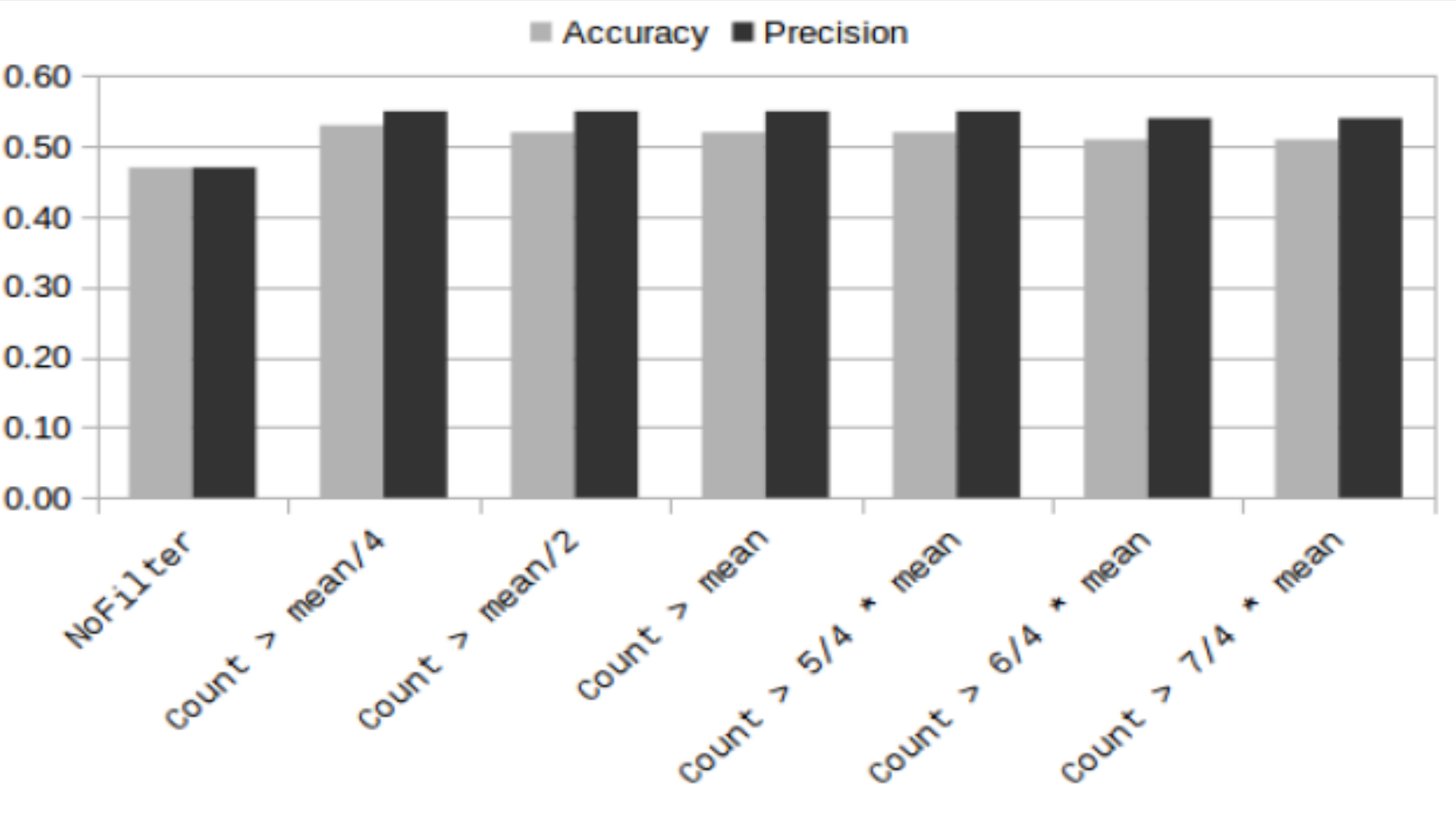
Build Filters

Range



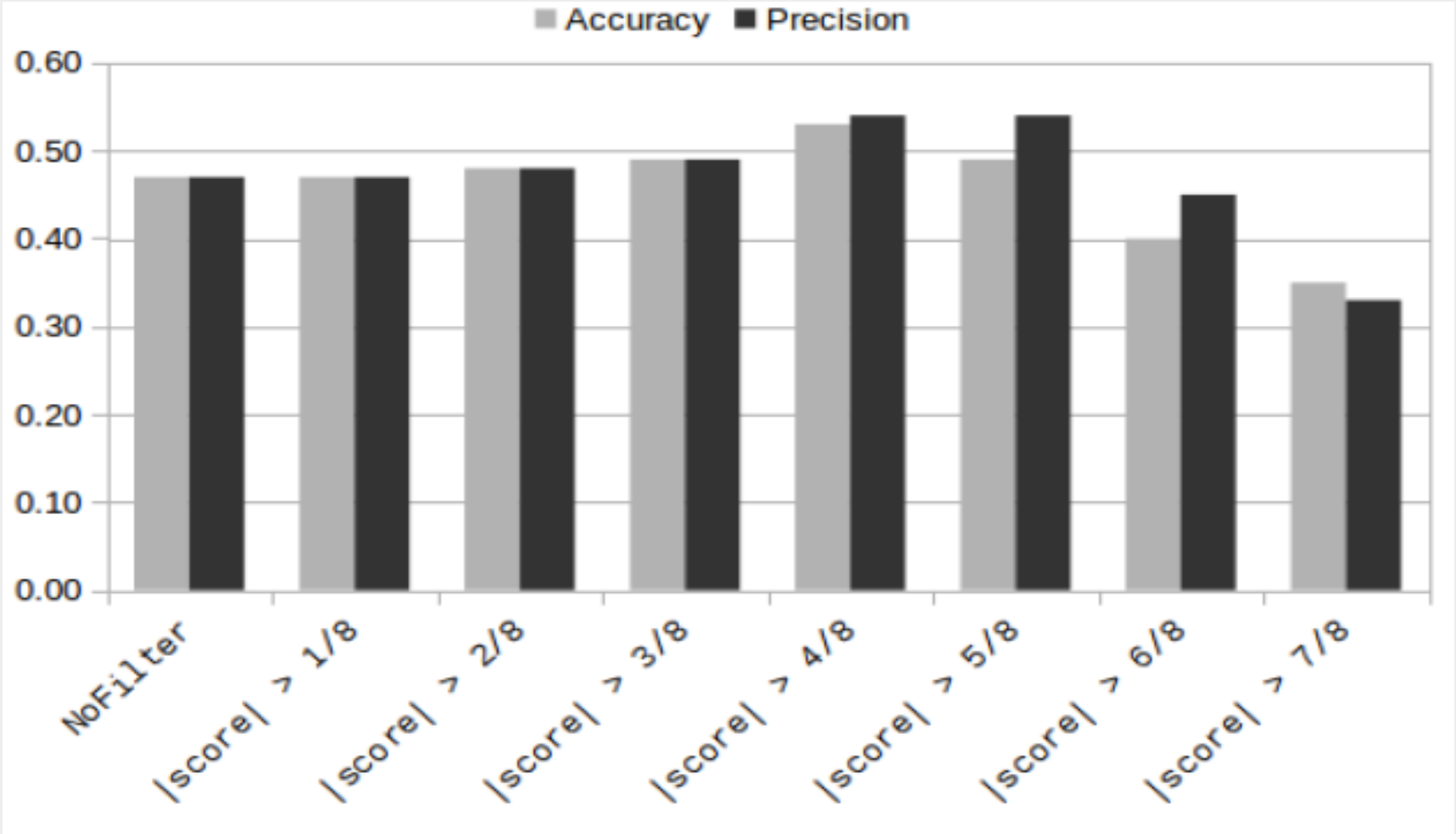
Build Filters

Count



Build Filters

SCORE



Unigrams (2-word Seed)

Unigrams (32-word Seed)

Unigrams (Emoticon Seed)

Bigrams (2-word Seed)

Bigrams (32-word Seed)

Bigrams (Emoticon Seed)