Corpus-based Construction of Polarity Lexicon to Identify Extreme Opinions

Sattam Almatarneh

Centro Singular de Investigación en Tecnoloxías da Información

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

citius.usc.es

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

CAMPUS VIDA

CiTiUS Centro Singular de Investigación en Tecnoloxías da Información

Outlines

1. **Introduction**
   - Objectives
   - Motivations
   - Sentiment Analysis(SA)
   - Sentiment Polarity Classification.
2. **Sentiment Lexicon Construction**
3. **Experiments and Evaluation**
4. **Conclusions and Discussion**

1. Introduction

❑ Objectives

- Propose a method for automatically building polarity lexicons from corpora

- Investigate the effectiveness of the automatic construction of a sentiment lexicon using unsupervised machine learning classification to search for extreme opinions.

CiTIUS

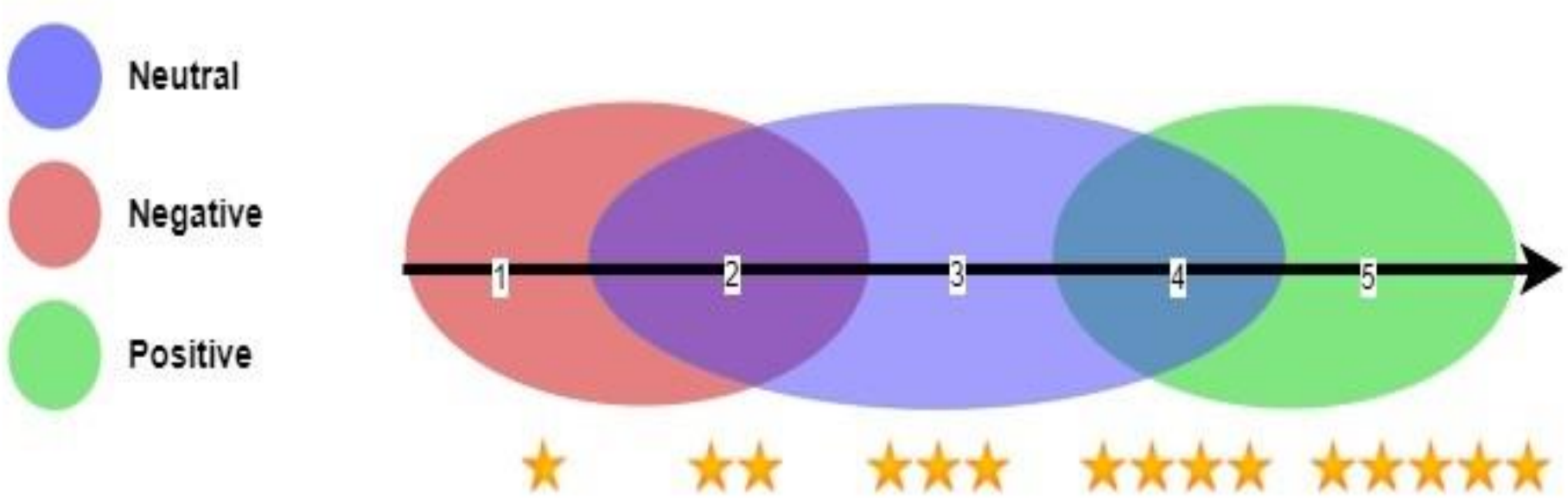
1. Introduction

❑ Motivations

- The Significant impact of Extreme Opinions in many fields such as industry, trade, political and social issues.
- Only about 5% of all opinions are in the most negative or the most positive level of the opinion scale,
- The construction of polarity lexicons is a strenuous and boring task if it is made manually.
- To the best of our knowledge, no sentiment analysis approach has considered the automatic identification and extraction of Extreme opinions

1. Introduction

Motivations



Hypothetical continuous distribution of negative, neutral and positive views on a scale from 1 to 5, according to the borderline between stars.

1. Introduction

❑ Sentiment Analysis(SA).

SENTIMENT ANALYSIS

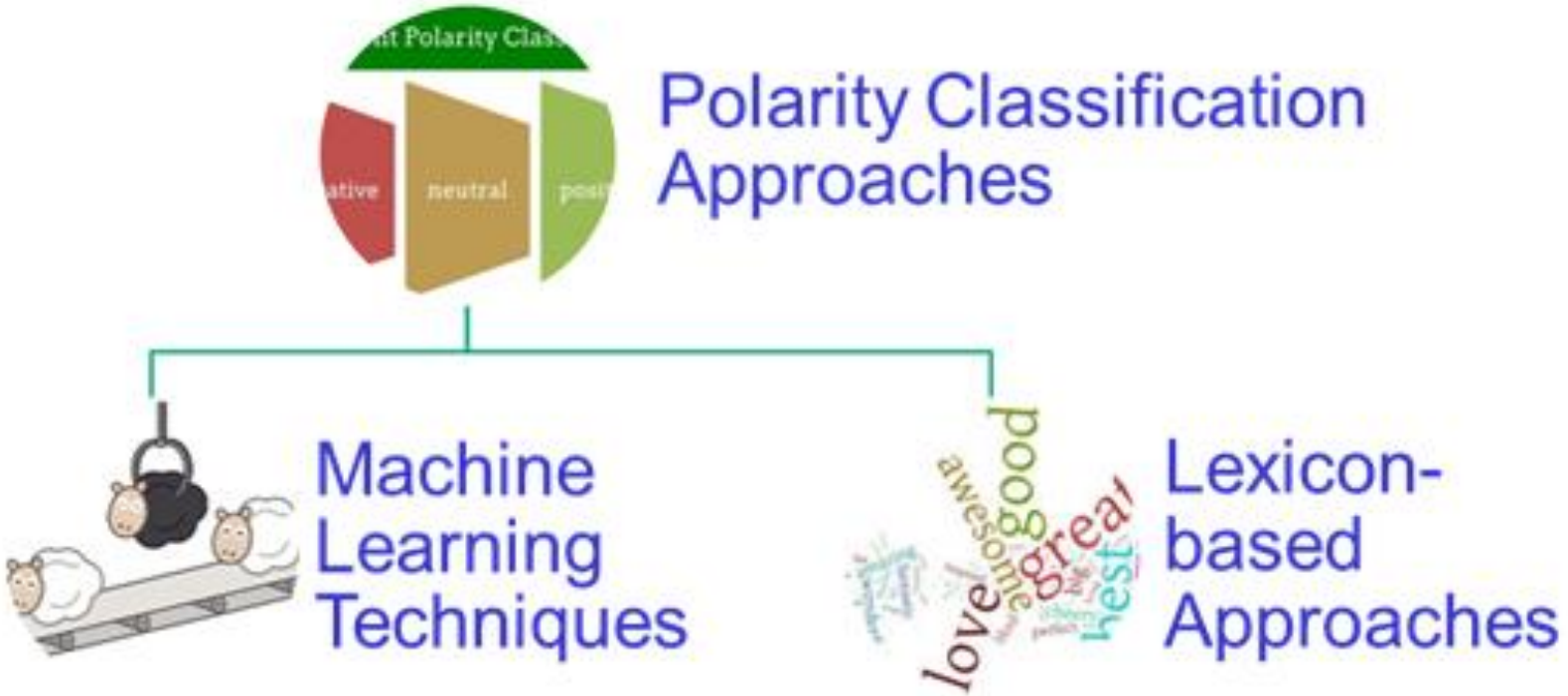
Discovering people opinions, emotions and feelings about a product or service

1. Introduction

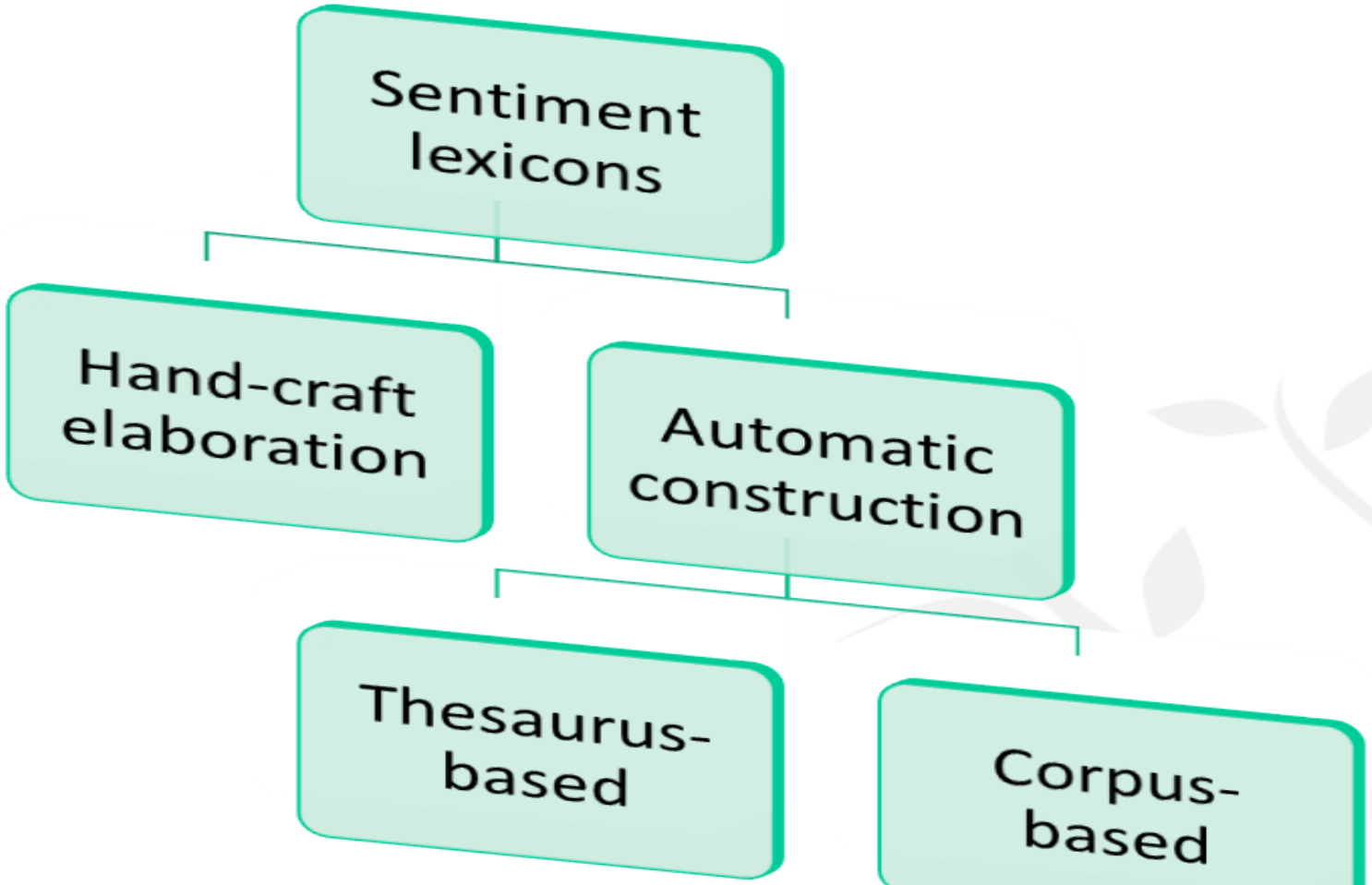
Sentiment Polarity Classification

1. Introduction

- **Sentiment Polarity Classification.**



2. Sentiment Lexicon Construction



2. Sentiment Lexicon Construction

- First step

$$RF_c(w) = \frac{freq(w,c)}{Total_c}$$

Word	Tag	Category	Freq	Total	Corpus
bad	a	1	122232	25395214	IMDB
bad	a	2	40491	11755132	IMDB
bad	a	3	37787	13995838	IMDB
bad	a	4	33070	14963866	IMDB
bad	a	5	39205	20390515	IMDB
bad	a	6	43101	27420036	IMDB
bad	a	7	46696	40192077	IMDB
bad	a	8	42228	48723444	IMDB
bad	a	9	29588	40277743	IMDB
bad	a	10	51778	73948447	IMDB

CiTIUS

2. Sentiment Lexicon Construction

- Second step

	Negatives				Neutral			Positive		
B=4	1	2	3	4	5	6	7	8	9	10
B=3	1	2	3	4	5	6	7	8	9	10
B=2	1	2	3	4	5	6	7	8	9	10
B=1	1	2	3	4	5	6	7	8	9	10

B=2	1	2	3	4	5
B=1	1	2	3	4	5

2. Sentiment Lexicon Construction

- Given a borderline value, *B*

$$AvMN(w) = \frac{\sum_{c=1}^{B} RF_c(w)}{B}$$

$$AvMP(w) = \frac{\sum_{c=(N+1)-B}^{N} RF_c(w)}{B}$$

$R = N - B$,where N is the total number of categories

$$AvNMN(w) = \frac{\sum_{c=B+1}^{N} RF_c(w)}{R}$$

$$AvNMP(w) = \frac{\sum_{c=1}^{N-B} RF_c(w)}{R}$$

CiTIUS

2. Sentiment Lexicon Construction

$$D_{neg}(w) = AvNMN(w) - AvMN(w)$$

IF $D_{MN}(w)$ < 0

IF $D_{MN}(w)$ > 0

1	2	3	4	5	6	7	8	9	10

$$D_{pos}(w) = AvNMP(w) - AvMP(w)$$

IF $D_{MP}(w)$ < 0

IF $D_{MP}(w)$ > 0

1	2	3	4	5	6	7	8	9	10

CiTIUS

3. Experiments and Evaluation

❑ Sentiment Classification

Algorithm 1 Identifying the most negative

```
docScore = 0
for Each document do
    WordScore = 0
    for Each Word in document do
        if Word IN MN then
            WordScore= -1 {the word in MN ( most negative words).}
        else
            WordScore= +1 {the word in NMN ( not most negative words).}
        end ifdocScore = docScore + WordScore
    end for
    if docScore <= 0 then
        docScore= Most Negative
    else
        docScore= Not Most Negative
    end if
    return docScore
end for
```

Algorithm 2 Identifying the most positive

```
docScore = 0
for Each document do
    WordScore = 0
    for Each Word in document do
        if Word IN MP then
            WordScore= +1 {the word in MP ( most positive words).}
        else
            WordScore= -1 {the word in NMP ( not most positive words).}
        end ifdocScore = docScore + WordScore
    end for
    if docScore >= 0 then
        docScore= Most positive
    else
        docScore= Not Most positive
    end if
    return docScore
end for
```

3. Experiments and Evaluation

Data Collection

Datasets	# of Reviews	MN	NMN	MP	NMP
Books	2000	522	1478	731	1269
DVDs	2000	530	1470	714	1286
Electronics	2000	666	1334	680	1320
Kitchens	2000	687	1313	754	1246
Movies	50000	14708	35292	14338	35662

Size of the five test datasets and the total number of reviews in each class the most negative Vs. Not most negative (MN vs. NMN) and the most positive Vs. not most positive (MP vs. NMP)

CiTIUS

3. Experiments and Evaluation

Lexicons

	Number of words			MN			NMN		
Lexicon	ADJ	ADV	Total	ADJ	ADV	Total	ADJ	ADV	Total
VERY-NEG B=1	**11670**	**2790**	**14460**	**4178**	**1092**	**5270**	**7492**	**1698**	**9190**
VERY-NEG B=2	**11557**	**2771**	**14328**	**4966**	**1266**	**6232**	**6591**	**1505**	**8096**
SO-CAL NP1	2826	876	3702	189	62	251	2637	814	3451
SO-CAL NP2	2826	876	3702	536	135	671	2290	741	3031
SO-CAL NP3	2826	876	3702	1080	289	1369	1746	587	2333
SO-CAL NP4	2826	876	3702	1576	429	2005	1250	447	1697
SentiWords NP1	13425	2811	16236	156	4	160	13269	2807	16076
SentiWords NP2	13425	2811	16236	1132	24	1156	12293	2787	15080
SentiWords NP3	13425	2811	16236	4016	189	4205	9409	2622	12031
SentiWords NP4	13425	2811	16236	7612	540	8152	5813	2271	8084

Negative lexicons: total number of words (adjectives and adverbs) for each lexicon, and number of words for each class (MP and NMP) in each lexicon

CiTIUS

3. Experiments and Evaluation

Polarity classification results for all collections with **SO-CAL ,SentiWords and VERY-NEG** lexicon, in terms of Precision (P_{neg}), Recall (R_{neg}) and $F1_{neg}$ scores for most negative (MN) and other (NMN) class of documents.

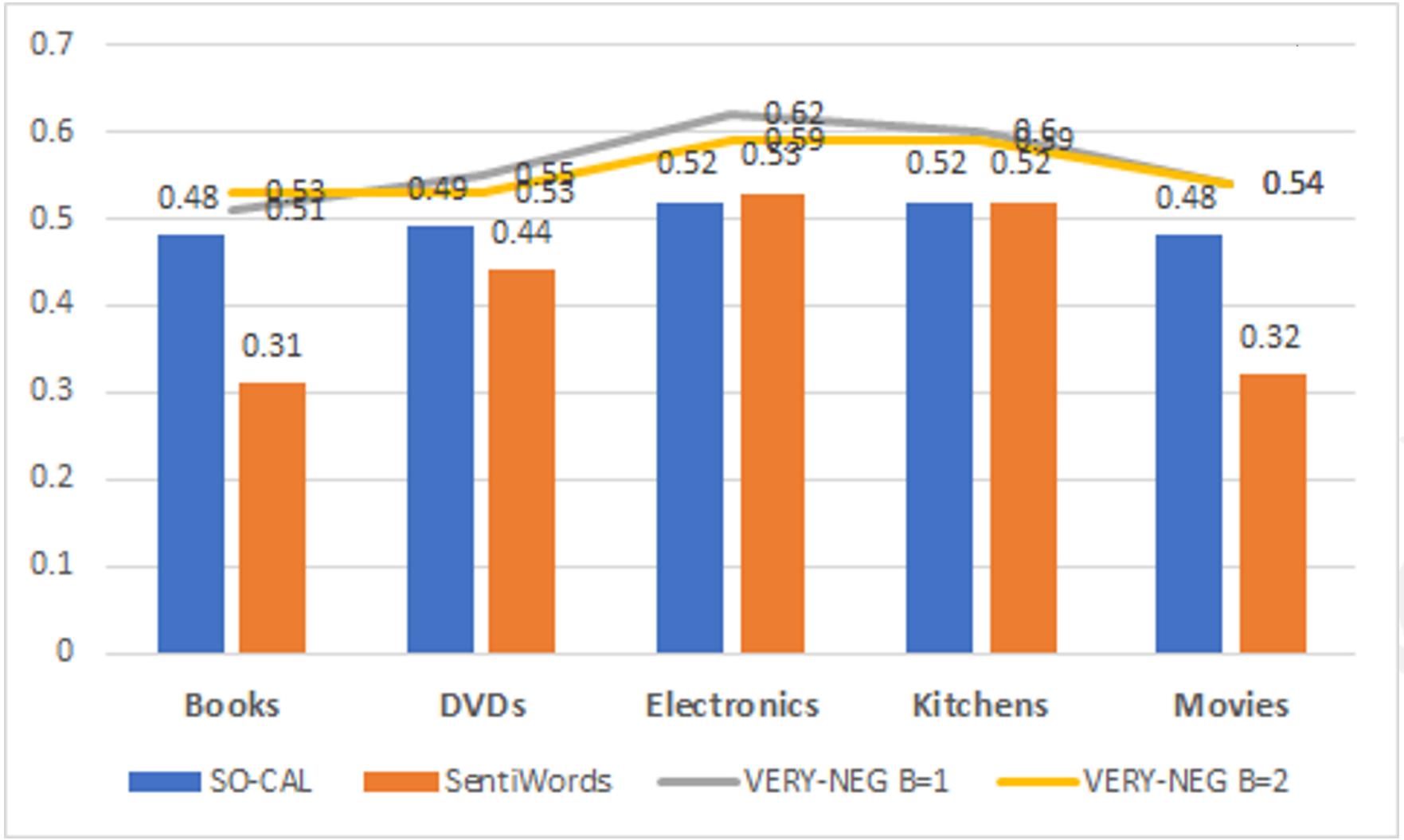
	NP1			NP2			NP3			NP4		
Dataset	P_{neg}	R_{neg}	$F1_{neg}$	P_{neg}	R_{neg}	$F1_{neg}$	P_{neg}	R_{neg}	$F1_{neg}$	P_{neg}	R_{neg}	$F1_{neg}$
Books	0.36	0.06	0.10	0.47	0.13	0.20	0.50	0.26	0.34	0.46	0.50	**0.48**
DVDs	0.60	0.10	0.17	0.58	0.18	0.28	0.56	0.31	0.40	0.48	0.51	**0.49**
Electronics	0.57	0.13	0.21	0.62	0.20	0.31	0.62	0.29	0.39	0.55	0.49	**0.52**
Kitchens	0.59	0.10	0.17	0.64	0.19	0.29	0.66	0.29	0.40	0.57	0.48	**0.52**
Movies	0.13	0.03	0.05	0.30	0.14	0.19	0.40	0.30	0.34	0.42	0.55	**0.48**

	NP1			NP2			NP3			NP4		
Dataset	P_{neg}	R_{neg}	$F1_{neg}$	P_{neg}	R_{neg}	$F1_{neg}$	P_{neg}	R_{neg}	$F1_{neg}$	P_{neg}	R_{neg}	$F1_{neg}$
Books	0.42	0.01	0.02	0.35	0.01	0.03	0.28	0.04	0.07	0.24	0.43	**0.31**
DVDs	0.33	0.01	0.01	0.53	0.03	0.06	0.58	0.13	0.22	0.49	0.41	**0.44**
Electronics	0.26	0.01	0.01	0.37	0.02	0.03	0.63	0.18	0.28	0.57	0.49	**0.53**
Kitchens	0.36	0.01	0.01	0.56	0.01	0.03	0.71	0.17	0.27	0.62	0.45	**0.52**
Movies	0.09	0.00	0.00	0.31	0.01	0.01	0.32	0.05	0.08	0.44	0.25	**0.32**

	VERY-NEG B=1			VERY-NEG B=2		
Dataset	P_{neg}	R_{neg}	$F1_{neg}$	P_{neg}	R_{neg}	$F1_{neg}$
Books	0.42	0.64	0.51	0.40	0.80	**0.53**
DVDs	0.43	0.76	**0.55**	0.88	0.88	0.53
Electronics	0.50	0.80	**0.62**	0.45	0.86	0.59
Kitchen	0.52	0.70	**0.60**	0.47	0.80	0.59
Movies	0.42	0.77	**0.54**	0.39	0.89	**0.54**

CiTIUS

3. Experiments and Evaluation



The best performance ($F1_{neg}$) obtained by all lexicons on all datasets for identifying most negative documents (MN vs NMN).

CiTIUS

3. Experiments and Evaluation

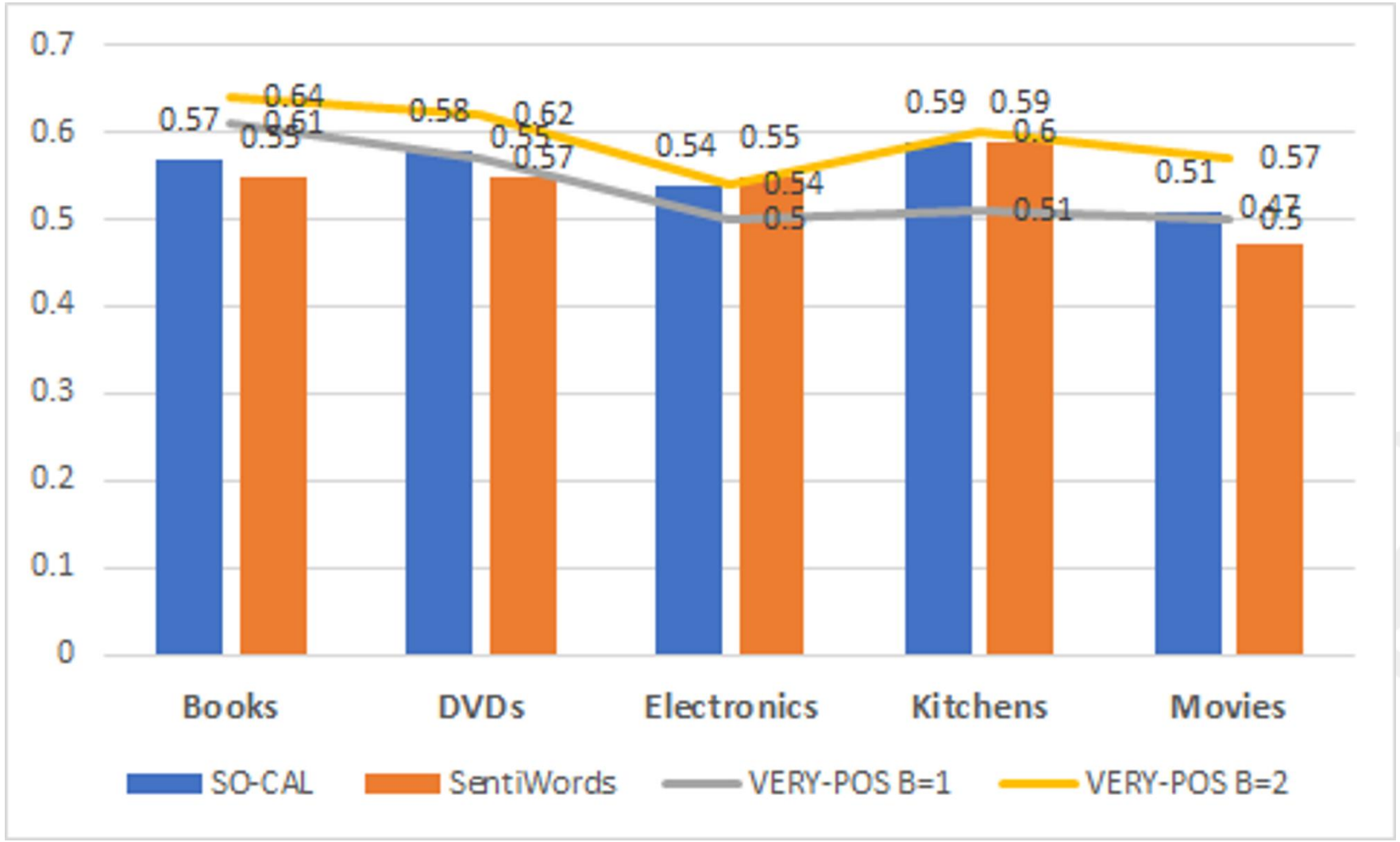
Polarity classification results for all collections with **SO-CAL** ,**SentiWords** and **VERY-POS** lexicon, in terms of Precision (P_{pos}), Recall (R_{pos}) and $F1_{pos}$ scores for most positive (MP) and other (NMP) class of documents.

	PP1			PP2			PP3			PP4		
Dataset	$\mathbf{P}_{pos}$	$\mathbf{R}_{pos}$	$\mathbf{F1}_{pos}$	$\mathbf{P}_{pos}$	$\mathbf{R}_{pos}$	$\mathbf{F1}_{pos}$	$\mathbf{P}_{pos}$	$\mathbf{R}_{pos}$	$\mathbf{F1}_{pos}$	$\mathbf{P}_{pos}$	$\mathbf{R}_{pos}$	$\mathbf{F1}_{pos}$
Books	0.61	0.17	0.27	0.54	0.34	0.42	0.52	0.55	0.53	0.41	0.94	**0.57**
DVDs	0.66	0.21	0.32	0.58	0.38	0.46	0.54	0.56	0.55	0.41	0.95	**0.58**
Electronics	0.54	0.26	0.35	0.51	0.40	0.45	0.49	0.60	0.54	0.38	0.94	**0.54**
Kitchens	0.53	0.23	0.32	0.53	0.36	0.43	0.50	0.55	0.52	0.42	0.97	**0.59**
Movies	0.75	0.11	0.20	0.60	0.29	0.39	0.52	0.49	0.50	0.35	0.94	**0.51**

	PP1			PP2			PP3			PP4		
Dataset	$\mathbf{P}_{pos}$	$\mathbf{R}_{pos}$	$\mathbf{F1}_{pos}$	$\mathbf{P}_{pos}$	$\mathbf{R}_{pos}$	$\mathbf{F1}_{pos}$	$\mathbf{P}_{pos}$	$\mathbf{R}_{pos}$	$\mathbf{F1}_{pos}$	$\mathbf{P}_{pos}$	$\mathbf{R}_{pos}$	$\mathbf{F1}_{pos}$
Books	0.76	0.06	0.12	0.66	0.13	0.22	0.60	0.38	0.46	0.40	0.93	**0.55**
DVDs	0.65	0.07	0.21	0.64	0.13	0.22	0.59	0.38	0.46	0.39	0.92	**0.55**
Electronics	0.70	0.11	0.19	0.71	0.19	0.30	0.63	0.41	0.50	0.40	0.93	**0.55**
Kitchens	0.61	0.07	0.13	0.63	0.17	0.27	0.65	0.37	0.47	0.43	0.94	**0.59**
Movies	0.64	0.01	0.03	0.63	0.05	0.09	0.55	0.27	0.36	0.31	0.95	**0.47**

	VERY-POS B=1			VERY-POS B=2		
Dataset	$\mathbf{P}_{pos}$	$\mathbf{R}_{pos}$	$\mathbf{F1}_{pos}$	$\mathbf{P}_{pos}$	$\mathbf{R}_{pos}$	$\mathbf{F1}_{pos}$
Books	0.67	0.55	0.61	0.61	0.67	**0.64**
DVDs	0.68	0.49	0.57	0.63	0.61	**0.62**
Electronics	0.63	0.42	0.50	0.57	0.52	**0.54**
Kitchen	0.63	0.43	0.51	0.60	0.60	**0.60**
Movies	0.63	0.41	0.50	0.55	0.58	**0.57**

3. Experiments and Evaluation



The best performance ($F1_{pos}$) obtained by all lexicons on all datasets for identifying the most positive documents.

CiTIUS

4. Conclusion and Discussion

❑ Conclusion

- **Method to automatically build a lexicon of extremely negative and positive words from labeled corpora.**
- **Put the stress on the extreme opinions because of their importance in various fields.**
- **Our classification algorithm is based on the very basic word-matching scheme to perform unsupervised sentiment analysis.**
- **Our automatically built lexicons have been fairly compared with handcrafted lexicons, by taking into account some partitions of them.**
- **The results of the experiments show that our lexicons are better suited for identifying the extreme opinions than two well-known resources: SO-CALL and SentiWords (a version of SentiWordNet).**

CiTIUS

4. Conclusion and Discussion

❑ Difficulties and challenges.

- **The borderline between very negative and not very negative or very positive and not very positive is still more difficult to find than that discriminating between positive and negative opinions.**

- **We made an exhaustive study of the effectiveness of linguistic features in supervised machine learning classification to search for the most negative opinions. The experiments we reported on that work showed low performance for all configuration systems. This means that the task of searching for extreme opinions is very challenging even for supervised strategies.**

Thank you for your attention

citius.usc.es