Multilingual Sentiment Analysis in Social Media

Supervisors

Dr. Rodrigo Agerri

Dr. German Rigau

Candidate

Iñaki San Vicente Roncal

March 11, 2019

Definition

Sentiment Analysis (SA) studies people's opinions, sentiments, and attitudes towards products, organizations, entities or topics.

Definition

Sentiment Analysis (SA) studies people's opinions, sentiments, and attitudes towards products, organizations, entities or topics.

WHY?

Definition

Sentiment Analysis (SA) studies people's opinions, sentiments, and attitudes towards products, organizations, entities or topics.

WHY?

- Organizations want to measure how the target consumers/social groups/audience react to their products/politics/proposals.
  - Surveys / Customer Services. $\rightarrow$ **Manual**, **great cost**, when feasible.
- Can we automatize the process? **WWW + NLP**

NLP challenges for SA

- Context dependent sentiment.

Example

"Gure salmentek behera egin dute"[a] vs. "Langabeziak behera egin du"[b]

[a] English: Our sales are going down.

[b] English: The unemployment rate is going down.

- Point of view

Example

"Osasunak 4-2 irabazi zuen Valladoliden aurka".[a]

[a] English: Osasuna won 4-2 against Valladolid.

NLP challenges for SA

- Sentiment granularity: document vs. phrases vs. words

Example

"Family hotel. Age is showing. Great$^{1.5}$ staff." A value hotel for sure with rooms that are average$^{-0.5}$, however some nice^{1} touches like the coffee station downstairs and the free^{1} brownies in the evening. Great$^{1.5}$ staff, super friendly^{2}. Special thanks to Camilla who was very helpful and forgiving, When we returned our damaged^{-1} umbrella.

- **Primary Goal: Develop Basque Sentiment Analysis**
- Is it enough to extract opinions exclusively in Basque?
  - Data is multilingual. Basque reality is multilingual (eu,es,fr).

- **Primary Goal: Develop Basque Sentiment Analysis**
- Is it enough to extract opinions exclusively in Basque?
  - Data is multilingual. Basque reality is multilingual (eu,es,fr).
- **Thesis Goal: Develop Multilingual Sentiment Analysis including Basque**

- Basque opinions in the web:
  - **Not supported**: TripAdvisor, Amazon, etc.
  - **Few specialized websites**, e.g., Armiarma (literature) or zinea.eus (movies).
  - Basque digital news media (Berria.eus, Sustatu.eus, Zuzeu.eus) **do not have active comment sections**.

Social Media

- Basque opinions in the web:
  - **Not supported**: TripAdvisor, Amazon, etc.
  - **Few specialized websites**, e.g., Armiarma (literature) or zinea.eus (movies).
  - Basque digital news media (Berria.eus, Sustatu.eus, Zuzeu.eus) **do not have active comment sections**.
- And Social Media?
  - **33.6%** of the population (16-50 year range, up to 80% of Twitter users) has activity in Basque (EAS).
  - **2.8 million tweets per year** in Basque (Umap)

Social Media: challenges

- Language identification

Example

"Kaixo, acabo de hacer la azterketa de gizarte. Fatal atera zait! 😭" [a]

[a] English: Hi, I just finished the exam of Social Studies class. I dit it awfully! :(

- Text normalization

Example

"Loo Exoo Maazooo dee Menooss Puuff :(" $\rightarrow$
"Lo hecho mazo de menos Puff :("[a]

[a] English: I miss him so much :(

Structure of this Thesis

Sentiment Lexicon Construction
- Subjectivity lexicons (Saralegi et al. , 2013) (CICLING)
- Automatic Sentiment lexicons (San Vicente et al. , 2014) (EACL)
- Method Comparison (San Vicente & Saralegi, 2016) (LREC)

Social Media Analysis
- Language Identification (Zubiaga et al. , 2016) (JLRE)
- Microtext Normalization (Alegria et al. , 2015; Saralegi & San Vicente, 2013) (JLRE)

Polarity Classification
- Spanish polarity Classification (San Vicente & Saralegi, 2014) (TASS)
- English polarity Classification (San Vicente et al. , 2015) (SemEval)

Real World Application
- Social Media Monitor (San Vicente et al. , 2019) (submitted to EAAI)
- Basque Polarity Classification

Conclusions
- Summary
- Future Work

Outline

Subjectivity Lexicons for less resourced languages (Saralegi et al. , 2013)

- **Compare methods for building sentiment lexicons:**
  - Projection/Translation (Mihalcea et al. , 2007)
  - Corpus-based lexicon generation (Turney & Littman, 2003)

- **Less resourced scenario:**
  - No use of MT systems.
  - No parallel corpora available.
  - No polarity annotated data-sets.

Projection/Translation

Approach

Translate an existing lexicon from other language by means of bilingual dictionaries.

- OpinionFinder (Wilson et al. , 2005) to Basque (en $\rightarrow$ eu)
- Only the first translation in $D_{en \rightarrow eu}$ (translations ordered by frequency of use).

Corpus-based Lexicon generation

Approach

Words that tend to appear in subjective (polar) texts with are good representatives of subjectivity (positive/negative polarity). $\rightarrow$ *Word Association measures*

- Log Likelihood Ratio (LLR) vs. Percentage Difference (%DIFF).

- **No corpus annotated with subjectivity!** $\rightarrow$ Heuristic:
  - Subjective: Opinion articles.
  - Objective: **Event news** vs. **Wikipedia**.

Subjective word distribution (Saralegi et al. , 2013)

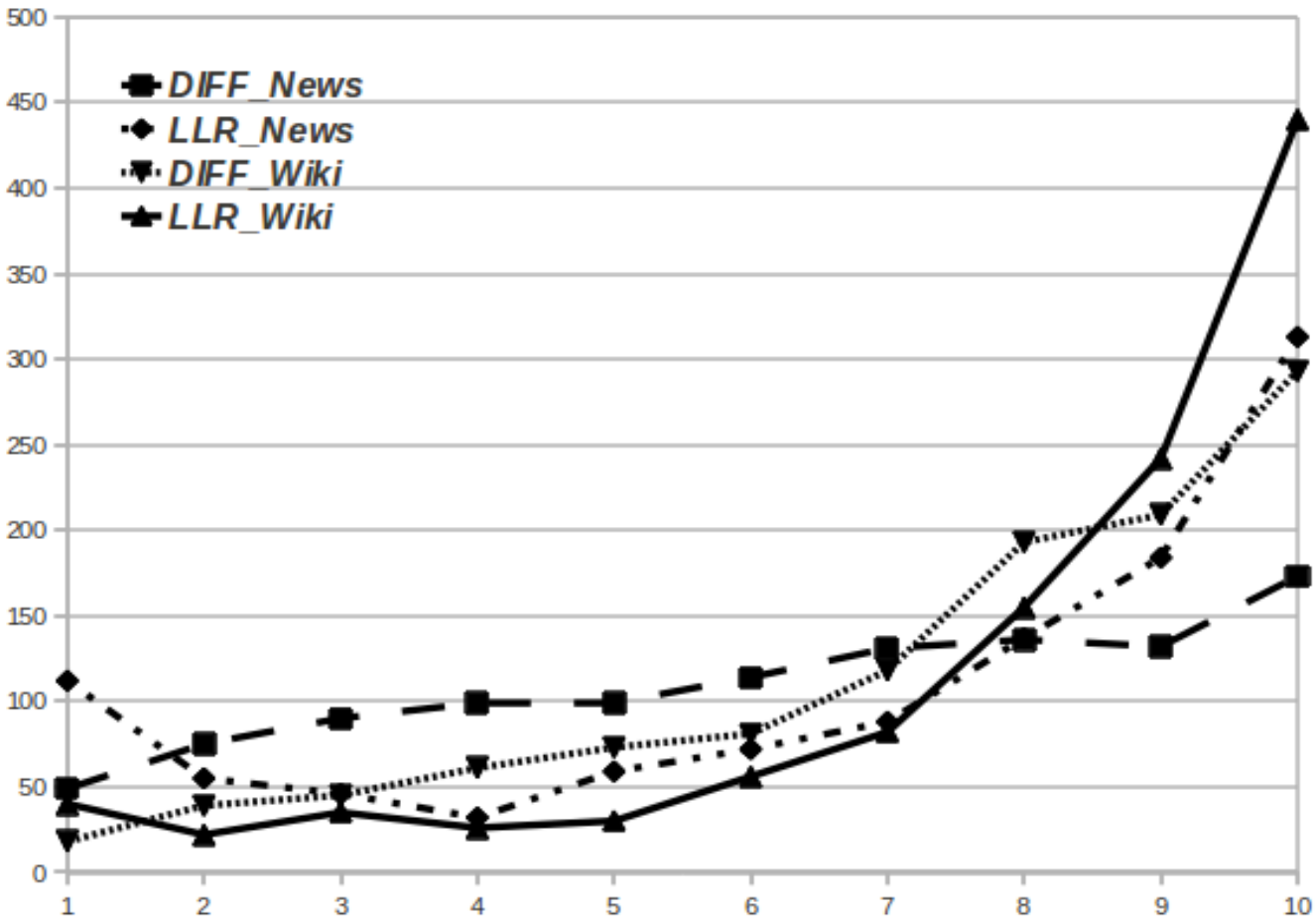


Figure – Distribution of subjective words with various measures and corpus combinations wrt. ranking intervals. Higher intervals contain words scoring higher in the rankings.

Subjectivity lexicons: evaluation (Saralegi et al., 2013)

- Subjectivity classification task.
- **New datasets in Basque**: 5 domains (journalism, blogs, Twitter, reviews, subtitles).
- Classifier:

$$subjectivity(tu) = \sum_{w \in tu} sub(w)/|tu| \quad (1)$$

Subjectivity lexicons: evaluation (Saralegi et al. , 2013)

- Subjectivity classification task.
- **New datasets in Basque**: 5 domains (journalism, blogs, Twitter, reviews, subtitles).
- Classifier:

$$subjectivity(tu) = \sum_{w \in tu} sub(w)/|tu| \quad (1)$$

- takeaways:
  - No lexicon is best :
    - Corpus based lexicons better for "in domain" (News)
    - Projection more robust across domains.
  - News better as objective corpus than Wikipedia.
  - LLR better than %DIFF for detecting subjective words.

Q-WordNet by Personalized Pageranking Vector (QWN-PPV)(San Vicente et al. , 2014)

Approach

Propagate the polarity of a few seeds through a Lexical Knowledge Base (LKB) projected over a graph

1. Seeds:
   - Synsets (Agerri & García-Serrano, 2010).
   - Words (Turney & Littman, 2003).
2. Propagation:
   - Graph: MCR (Agirre et al. , 2012).
   - Algorithm: UKB Personalized PageRank propagation algorithm (Agirre & Soroa, 2009):

$$\mathbf{Pr} = cM\mathbf{Pr} + (1-c)\mathbf{v}$$

QWN-PPV: Evaluation (San Vicente et al. , 2014)

- Task based evaluation: polarity classification.
  - **3 datasets**: MPQA (en), (Bespalov et al. , 2011) (en), HOpinion (es).
  - **7 sentiment lexicons**:
    - Automatic={SWN, MSOL, QWN}
    - (semi-)Manual={Liu, GI, SO-CAL, OF}
  - Classifier:

$$polarity(d) = \frac{\sum_{w \in d} pol(w)}{|d|} \tag{2}$$

QWN-PPV: Evaluation (San Vicente et al. , 2014)

- Task based evaluation: polarity classification.
  - **3 datasets**: MPQA (en), (Bespalov et al. , 2011) (en), HOpinion (es).
  - **7 sentiment lexicons**:
    - Automatic={SWN, MSOL, QWN}
    - (semi-)Manual={Liu, GI, SO-CAL, OF}
  - Classifier:

$$polarity(d) = \frac{\sum_{w \in d} pol(w)}{|d|} \tag{2}$$

- takeaways:
  - No lexicon is best throughout all datasets $\rightarrow$ QWN-PPV produces task specific lexicons.
  - **Outperforms** automatic methods, **competitive** vs. manual lexicons.
  - Only needs a Wordnet like LKB.

Comparing methods: Basque (San Vicente & Saralegi, 2016)

- Objectives:
  - compare the previous approaches.
  - Generate the polarity lexicons for Basque.

- When facing the task of creating such a resource for a new language:
  - **Is it worth to make a great manual annotation effort?**

Lexicons generated (San Vicente & Saralegi, 2016)

Lexicon	#Lemmas	#+ lemmas	#- lemmas	Annotation speed	Annotation time (h)
Lex_{pr}	5.335	1.892	3.303	5.3 w/min	36h
Lex_{C}	1.660	959	691	8.3 w/min	10h
$Lex_{Qwn-ppv}$	1.132	565	567	-	-

Table – Statistics for the lexicons generated.

- Projection Lex_{pr}: $ElhPolar_{es}$ (Saralegi & San Vicente, 2013) $\rightarrow$ eu. 5 translations per entry.
- Corpus-based Lex_{C}: subjective/objective corpus (Saralegi et al., 2013) + positive/negative manual annotation (5.000).
- Automatic $Lex_{Qwn-ppv}$: MCR synonym/antonym graphs. Setup from (San Vicente et al., 2014) experiments.

Manual Effort: Projection vs. Corpus-based (San Vicente & Saralegi, 2016)

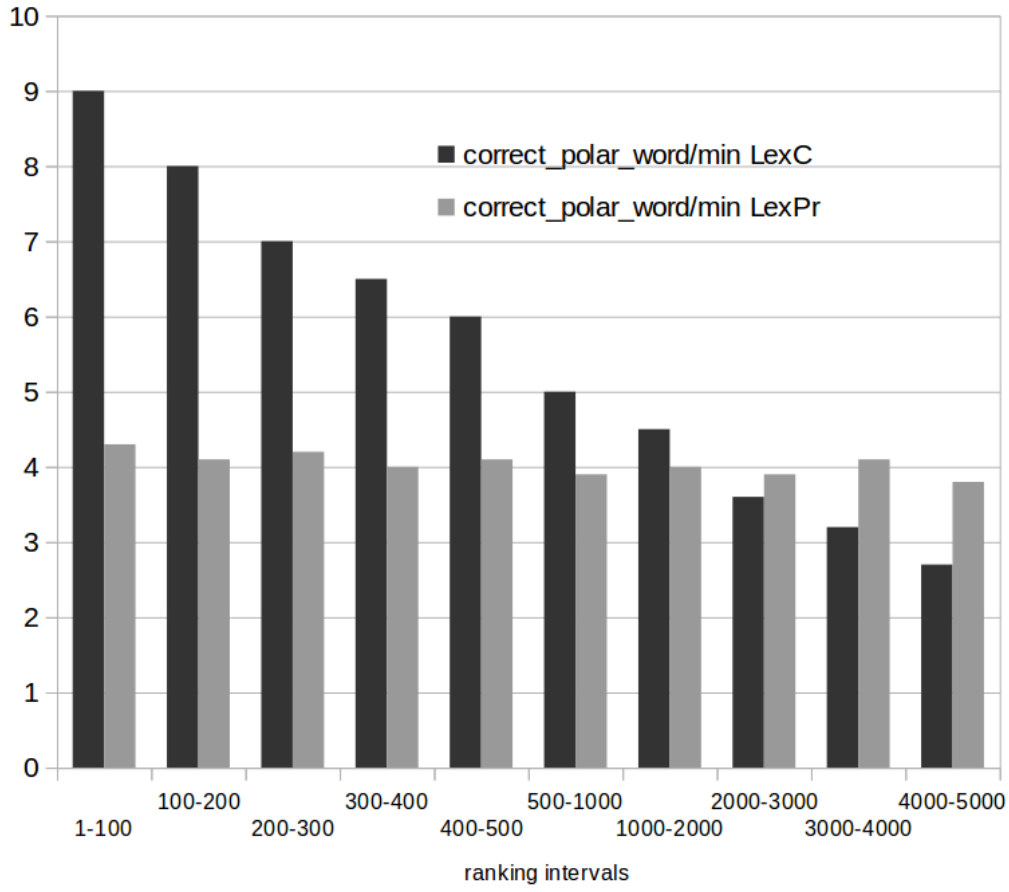


Figure – Correction speed and productivity data for Lex_{pr} and Lex_c.

Results for Basque (San Vicente & Saralegi, 2016)

Lexicon	News			Music&Films			Overall		
	Acc.	Fpos	Fneg	Acc.	Fpos	Fneg	Acc.	Fpos	Fneg
Projection									
Lex_{pr}	**0.86**	0.68	0.91	0.70	0.75	0.62	**0.79**	0.72	0.84
Corpus-based									
Lex_{c}	**0.78**	0.56	0.86	**0.80**	0.86	0.67	**0.79**	0.75	0.82
Automatic									
$Lex_{qwn-ppv}$	0.67	0.21	0.79	0.55	0.68	0.20	0.63	0.53	0.69
Combination									
ConsensLex_{c+pr}	**0.88**	0.74	0.92	**0.83**	0.87	0.73	**0.86**	0.82	0.88
External									
NRC_{eu}	0.62	0.29	0.74	0.47	0.51	0.41	0.56	0.41	0.65
MLSenticon	0.65	0.37	0.76	0.55	0.60	0.48	0.61	0.50	0.68

Table – Projection > Corpus-based > LKB-based.

Results for Basque (San Vicente & Saralegi, 2016)

Lexicon	News			Music&Films			Overall		
	Acc.	Fpos	Fneg	Acc.	Fpos	Fneg	Acc.	Fpos	Fneg
Projection									
Lex_{pr}	**0.86**	0.68	0.91	0.70	0.75	0.62	**0.79**	0.72	0.84
Corpus-based									
Lex_{c}	**0.78**	0.56	0.86	**0.80**	0.86	0.67	**0.79**	0.75	0.82
Automatic									
$Lex_{qwn-ppv}$	0.67	0.21	0.79	0.55	0.68	0.20	0.63	0.53	0.69
Combination									
Lex_{c+pr}	**0.88**	0.74	0.92	**0.83**	0.87	0.73	**0.86**	0.82	0.88
External									
NRC_{eu}	0.62	0.29	0.74	0.47	0.51	0.41	0.56	0.41	0.65
MLSenticon	0.65	0.37	0.76	0.55	0.60	0.48	0.61	0.50	0.68

Table – QWN-PPV better than other external lexicons.

Contribution table

Publication	Topic(s)	Langs	Datasets	Resources	Software
(Saralegi et al. , 2013)	Subjectivity Lexicons - Translation, Corpus based	Eu	News, blogs, tweets, Music/Film reviews	Lexicons (eu, corpus based and translated)	DSPL
(San Vicente et al. , 2014)	Sentiment Lexicons - LKB based	En, Es	-(Bespalov et al. , 2011)* -MPQA* -HOpinion*	-Lexicons (es,en)	QWN-PPV
(San Vicente & Saralegi, 2016)	Sentiment Lexicons - comparison	Eu	News, Music/Film reviews	-*ElhPolar*$_{eu}$ lexicon -QWN-PPV lexicons for Basque	-

- First subjectivity and sentiment lexicons for Basque.
- Task based (extrinsic) evaluations.
- Publicly available software.

Outline

TweetLID Shared task (Zubiaga et al. , 2016)

- Goal: **Identify language of tweets** - (ca,es,eu,gl,pt) + English

Example

Qeeeee matadaaa [a] *da Biyar laneaaaa...* [b] $\rightarrow$ *es+eu*

[a] **English: that was exhausting (es)**

[b] **English: and gotta go to work tomorrow (eu)**

- 7 participants, 21 systems
- Benchmark for LID focused on less-resourced languages
- **Role** as organizer: Annotation, coordination, evaluation.

TweetLID: Datasets

- 35k Tweets (Train 15K / Test 20K) fitting geographical criteria:
  - **Portugal**.
  - **Basque Country**, where Basque and Spanish are spoken → Gipuzkoa
  - **Catalonia**, where Catalan and Spanish are spoken → Girona
  - **Galicia**, where Galician and Spanish are spoken → Lugo

- Multi-label annotation:
  - Ambiguous tweets: e.g. *Acabo de publicar una foto* [a] → *ca/es*.
  - Multilingual tweets.

[a] **English: I just published a photo**

TweetLID: Results per language

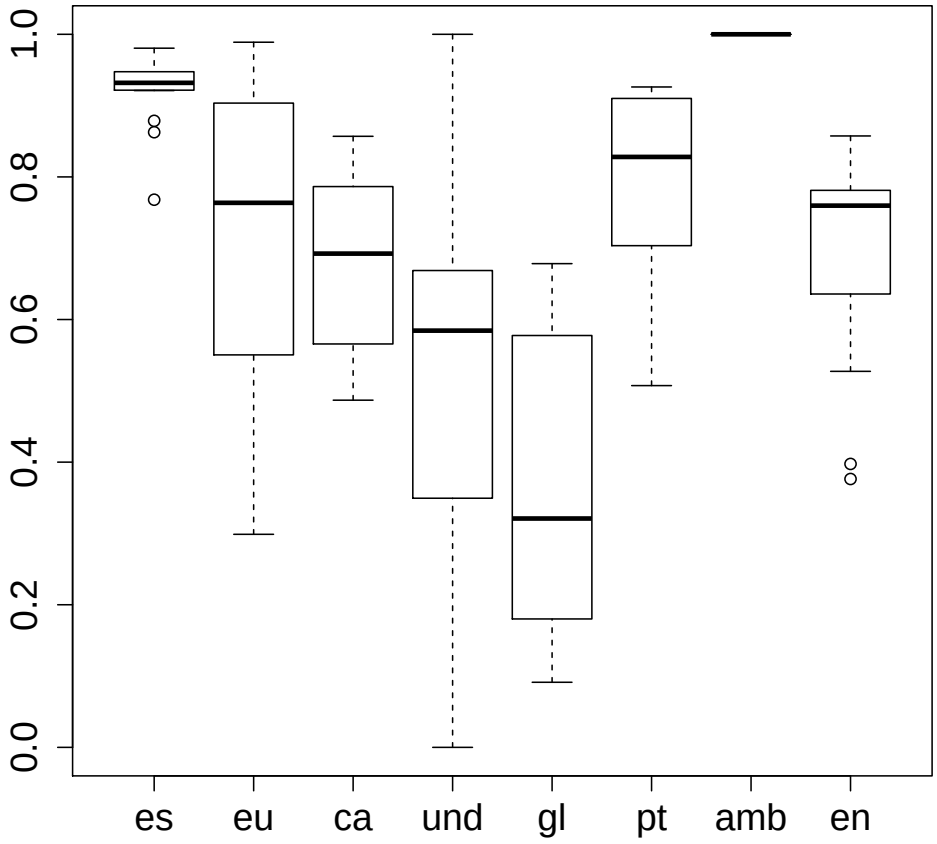


Figure – Distribution of precision scores by language for the 21 submitted systems, including results for both the constrained and the unconstrained tracks.

TweetLID: Takeaways

- Word and character ngrams used.
- **Normalization**: remove URL, @, #, uppercase,repeated characters.
- External resources **not** useful.
- Best **Microavg. Acc. 89.9%** (Macroavg Acc. 82.5%). State of the art (major languages): 92,4% (Carter et al. , 2013)
- Short tweets are difficult (<60 chars).
- **Multilingual tweets pending** (2/7 participants).

TweetNorm Shared Task (Alegria et al. , 2015)

- Goal: Normalization of Tweets in Spanish

Example

cariiii k no te seguia en twitter!!!mu fuerte!!!...se te exa d menos en el bk....sobreto en los cierres jajajajas[a]

→

cariño que no te seguía en twitter!!!muy fuerte!!!...se te echa de menos en el **bk**....sobre todo en los cierres ja

[a] **English: my dear i wasn't following you on twitter!!no way!! we miss you in the bk.... especially when closing hahaha**

- 13 participants
- Benchmark for Microtext Normalization
- **Role** as organizer: coordination, evaluation.

TweetNorm: Elhuyar (Saralegi & San-Vicente, 2013)

- Two step algorithm:
  1. Generates all the possible candidates for the OOV words in a tweet.
     - Rules, LCSR: common abbreviations, colloquial expressions, repeated characters, onomatopoeia and orthographic errors.
     - Reference lexica of normalized forms were generated from various resources.
  2. Selects the combination of candidates that best fits a LM.
     - SRILM based on bigrams obtained from Wikipedia articles and a news corpus from EFE.
- ranked 4th.
- To improve: OOVs containing several errors.

Example

'cumpleee' $\rightarrow$ 'cumple' $\rightarrow$ 'cumpleaños'

TweetNorm: Results

Rank	System	Prec1	Prec2
—	*Oracle*	0.927	—
1	RAE	0.781	—
2	Citius-Imaxin	0.663	0.662
3	UPC	0.653	—
4	Elhuyar	0.636	0.634
5	EHU	0.619	0.609
...	...	...	...
—	*Baseline*	0.198	—

- Generate/Filter strategy: 10 out of 13 systems.
- Generate: Rules, RE, transducers, edit distance, gazetteers.
- Filter: LM (1-5grams), scoring.

Contribution table

Publication	Topic(s)	Langs	Task	Datasets	Resources	Software
(Zubiaga et al. , 2016)	Language identification in Twitter	Ca, Gl, En, Es, Eu, Pt	TweetLID	TweetLID corpus	-	-
(Alegria et al. , 2015)	Microtext Normalization	Es	TweetNorm	TweetNorm corpus	-	-
(Saralegi & San Vicente, 2013)	Microtext Normalization	Es	TweetNorm	TweetNorm corpus*	OOV normalization dictionary (es)	Normalization module

- Organizer of TweeLID and TweetNorm shared tasks.
- Generated benchmarking datasets.
- TweetNorm participation $\rightarrow$ Normalization module.

Outline

Spanish Polarity Classification

- 3 participations in TASS (2012, 2013, 2014)
- (Saralegi & San Vicente, 2012) (rank: 1st)
  - *ElhPolar*$_{es}$ v1. Projection + corpus-based.
  - ngrams vs. **Polarity lexicon lemmas.**
  - Twitter normalization: Emoticons, interjections, urls.
- (Saralegi & San Vicente, 2013) (rank: 1st)
  - *ElhPolar*$_{es}$ v2.
  - TweetNorm normalization (Saralegi & San Vicente, 2013)
  - Polarity scores based on *ElhPolar*$_{es}$ include **modifiers**.
- (San Vicente & Saralegi, 2014) (rank: 2nd)
  - Syntax based ngrams. E.g. *perro faldero* [Noun+Adj]
  - Negation treatment features: *w* and *NOT_w*
  - **Lexicon Combination.**

TASS Takeaways

- 👍:
  - *ElhPolar*$_{es}$ key to success.
  - Polarity scores.
  - Normalization helps.
- 👎:
  - Additional training examples.
  - performance of NEU.
  - Train/test corpora distribution.

English Polarity Classification (San Vicente et al. , 2015)

- Semeval 2015 ABSA shared task.
  - Domains: Restaurant, Laptops, Hotels (no training data)
- Features different wrt. the Spanish system:
  - Domain specific sentiment lexicons (Yelp, Amazon).
  - Word Clusters (word2vec + K-means) from Yelp, Amazon.
  - Category information (present in the datasets).

SemEval Results (EN) (San Vicente et al. , 2015)

System	Rest.	Lapt.	Hotel
Baseline	63.55	69.97	71.68 (majority)
Sentiue	*78.70 (1)*	*79.35 (1)*	71.68 (4)
lsislif	75.50 (3)	77.87 (3)	*85.84 (1)*
EliXa (u)	70.06(10)	72.92 (7)	79.65 (3)
EliXa (c)	67.34 (14)	71.55 (9)	74.93 (5)

Table – Results obtained on the slot3 evaluation on restaurant data; ranking in brackets.

- takeaways:
  - **ngrams** vs. polarity lexicon ngrams.
  - Domain polarity lexicons.
  - Clusters need lots of data.

EliXa

- http://github.com/Elhuyar/Elixa
- **SVM + linguistic features**:
  - word form/ lemma n-grams.
  - PoS tags.
  - **Sentiment lexicon lemmas/ polarity scores**.
  - Polarity modifiers (good $\neq$ not good $\neq$ very good).
  - Interjections, onomatopoeia.
  - Typographic polarity clues: punctuation, uppercase.
  - Cluster features.
- 4 languages: EU,EN,ES,FR
- **Ixa-pipes integrated**
- Inherent problems of social media addressed $\rightarrow$ **Microtext normalization**
  - Non standard language, emojis (Saralegi & San Vicente, 2013) $\rightarrow$ **SA oriented**.

Contributions in polarity classification

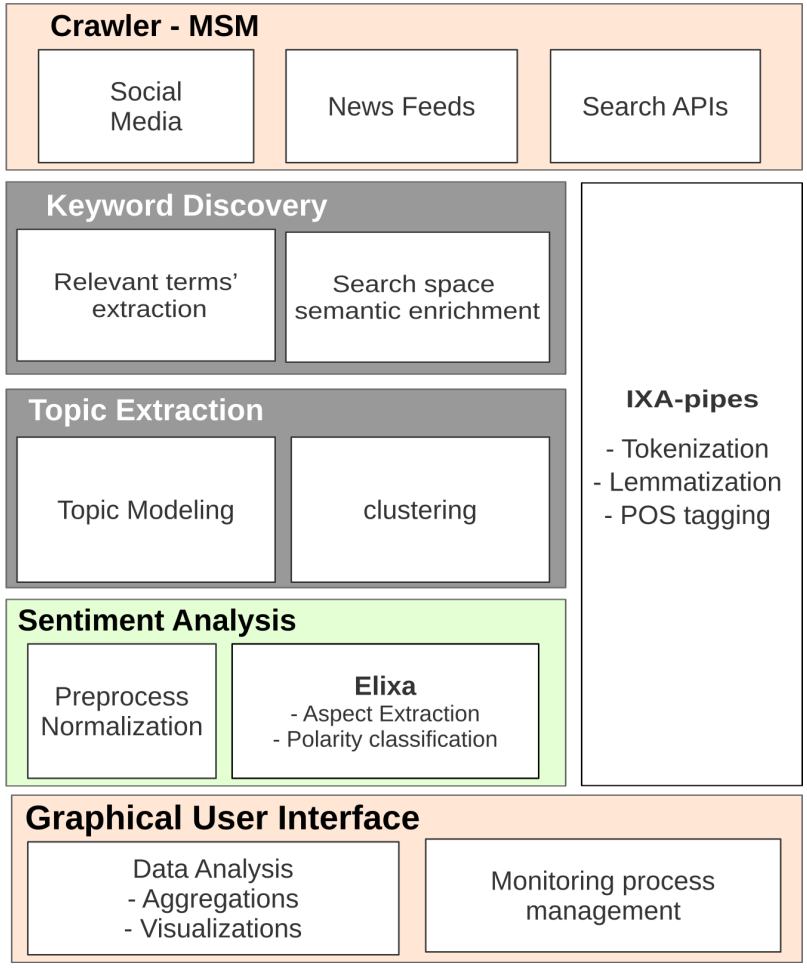
Publication	Topic(s)	Langs	Task	Resources	Software
(San Vicente & Saralegi, 2014)	Polarity classification	Es	TASS	*ElhPolar*$_{es}$ lexicon	SVM classifier
(San Vicente et al., 2015)	Polarity classification, Aspect Based SA	En	SemEval ABSA	Sentiment Lexicons (en, domain specific)	EliXa

- Sentence and document level polarity classification.
- 3 participations in TASS (es): **1st**(2012), **1st**(2013), **2nd** (2014)
- SemEval ABSA 2015. **3rd** in hidden domain task.
- First release of EliXa SA software, **open source**.

Outline

What is Talaia?

Automatic analysis of the impact in social media and digital press of topics specified by the user, based on Natural Language Processing.



Talaia: Success cases: Behagunea

- Real time opinion monitor - Donostia 2016 cultural capital
  - Basque, English, French, Spanish.
  - Developed by **Elhuyar and IXA**. Competitive tendering.
  - Low latency: **166K mentions in a year (max 6.6K mentions/day)**.
  - Real time: 15 minutes.
- http:/behagune.elhuyar.eus

Talaia: Success cases: Basque elections 2016

- Real time opinion monitor - Basque regional election campaign 2016.
  - Basque, Spanish.
  - Limited geographical area.
  - Collaboration with Berria.
  - Data volume: 4.25M mentions (avg. 125K mentions/day, **max. 433K mentions/day**).
- http://talaia.elhuyar.eus/demo_eae2016

Talaia: Datasets

- No datasets for training supervised systems. **Two new multilingual datasets** created:

Language	Total size	#pos	#neg	#neu
eu	2937	931	408	1598
es	4754	1487	1303	1964
en	12,273	4,654	1,837	5,782
fr	11,071	3,459	2,618	4,994

Table – Cultural domain dataset in Basque.

Language	#Tweets	#Annotations	#pos	#neg	#neu
eu	9,418	11,692	3,974	3,185	4,533
es	15,550	20,278	3,788	7,601	8,889

Table – Political domain dataset in Basque, entity level annotations.

Talaia: Results

Language	#features	acc	fpos	fneg	fneu
Cultural Domain					
eu	4,777	74.02	0.658	0.635	0.803
es	10,037	73.03	0.683	0.756	0.744
en	24,183	70.43	0.715	0.530	0.743
fr	23,779	66.17	0.600	0.617	0.721
Political Domain					
eu	9,394	69.88	0.714	0.702	0.683
es	15,751	67.05	0.545	0.693	0.700

- SVM Features:
  - 1-gram word forms (frequency $>= 2$; document frequency (df) $>= 2$).
  - POS tag 1-gram features.
  - Polarity lemmas in $ElhPolar_{eu}$ (San Vicente & Saralegi, 2016).
  - Sentence length.
  - Upper case ratio: % of capital letters wrt. sentence length in characters.

Contribution table

Publication	Topic(s)	Langs	Datasets	Resources	Software
(San Vicente et al. , 2019)	Social Media monitor, normalization, Polarity classification	En, Es, Eu, Fr	-DSS2016 Behagunea -BEC2016 (politics)	Social media normalization resources	-Behagunea UI -MSM crawler -EliXa

- Integration of previous research.
- First full SA system including Basque.
- First polarity annotated datasets for Basque.
- System in production.
- Open source software.

Outline

Summary

- Multilingual Sentiment Analysis in order to develop a social media monitor on specific topics, including Basque.
  - Methods applicable across languages.
  - Methods applicable to less-resourced languages.

Summary: Sentiment Lexicons

- Pioneering work for Basque:
  - First sentiment lexicons (subjectivity/polarity).
- Novel method for automatic lexicon construction. Publicly available https://github.com/ixa-ehu/qwn-ppv
- **Evaluation** of Sentiment lexicons must be **task-based**.
- For Basque manual effort pays off vs. fully automatic methods(San Vicente & Saralegi, 2016).

Summary: Social Media

- Part of the organizing committee in two shared tasks:
  - TweetLID: Annotation, coordination, evaluation.
  - TweetNorm: coordination, evaluation.
- Participant in TweetNorm (ranked **4th**).
- Multi Source Monitor (MSM): Publicly available software to harvest data from social Media (Twitter) (San Vicente et al. , 2019).
  https://github.com/elhuyar/MSM
- Pending issues:
  - Identification of multilingual tweets and short messages (<60 chars).
  - Task dependent normalization.

Summary: Polarity Classification

- Pioneering work for Basque:
  - The first polarity annotated datasets.
    https://hizkuntzateknologiak.elhuyar.eus/eu/baliabideak
  - We generated the first resources for Basque microtext normalization
    https://hizkuntzateknologiak.elhuyar.eus/assets/files/elixa-resources-10.tgz
  - EliXa, the first multilingual SA system including Basque
    https://github.com/elhuyar/elixa
- Participation in international shared tasks:
  - TASS (es): **1st**(2012), **1st**(2013), **2nd** (2014)
  - SemEval ABSA 2015 (en). **3rd** in hidden domain task.
- Pending: aspect extraction

Summary: Real World application

- Talaia https://talaia.elhuyar.eus
  - Culmination of the journey → Final product
  - System in production.
  - **Open source** software.

Summary: Thesis in Numbers

- 14 publications.
- 2 shared tasks organized.
- 5 participations in shared tasks.
- 5 software packages publicly available.
- 1 final product.
- Previously non existing SA resources for Basque:
  - Polarity lexicons for Basque (2+).
  - 2 Polarity annotated datasets.

Future Work

- Polarity classification:
  - Deep EliXa:
    - Robust cross domain performance
    - Cost of training and hyper-parameter tuning vs. improvement obtained over other approaches.
  - Domain adaptation: measure the cost of creating datasets for new domains.
- Aspect Based Sentiment Analysis
- Data crawling
  - Keyword based crawling suffers from coverage, keywords change over time.

Acknowledgements

Projects

ber2tek

OpeNER

ElkarOla

TUNER

Tacardi
Traduccion automática en contexto y aumentada con recursos dinámicos de internet

Knowtour
(IE11-305)

skater

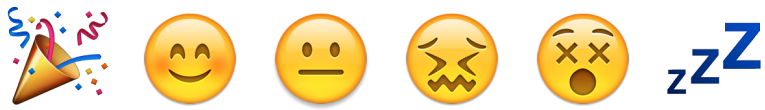
NewsReader

Organizations

elhuyar

eman ta zabal zazu
Universidad del País Vasco
Euskal Herriko Unibertsitatea

ixa



Eskerrik asko!

Moltes gràcies!

Thank you!

¡Muchas gracias!