TASS 2012

Lexicon-Based Sentiment Analysis of Twitter Messages in Spanish

Antonio Moreno Ortiz
Chantal Pérez Hernández
Universidad de Málaga
{amo, mph}@uma.es

Machine Learning vs. Lexicon-based SA

Machine Learning	Lexicon-based
Requires a training data set	No training set required
Output is a classification	Output is a numerical rating on a scale
Good results for analysis data similar to training data	Reasonable results across any texts
Better at domain-specific texts (for those domains they are trained!)	Better at general language texts. Harder to integrate domain-specific knowledge
Faster to develop from readily available ML software	Construction of linguistic resources is time consuming

Sentitext

- Client-server architecture
- Knowledge / processor independence
- Knowledge sources:
  - Individual Words Lexicon
  - Multiword Lexicon
  - Context Rules (Contextual Valence Shifters)
- Uses Freeling for lemmatization and phrase identification
- Returns a Global Sentiment Value (GSV) for the whole input text, plus identified individual sentiment-carrying text segments
- Semantic disambiguation only through CVS

Context Rules (CVS)

- Our implementation of Contextual Valence Shifters (Polanyi & Zaenen)
- Stored as a lexical resource (i.e., not hard-coded).
- 3 types:
  - Intensifying: e.g. “muy + ADJ” = INT1
  - Downtoning: e.g. “un poco + ADJ” = DOW1
  - Reversing: “no + (ser) + nada + ADJ” = +/-X

Global Sentiment Value

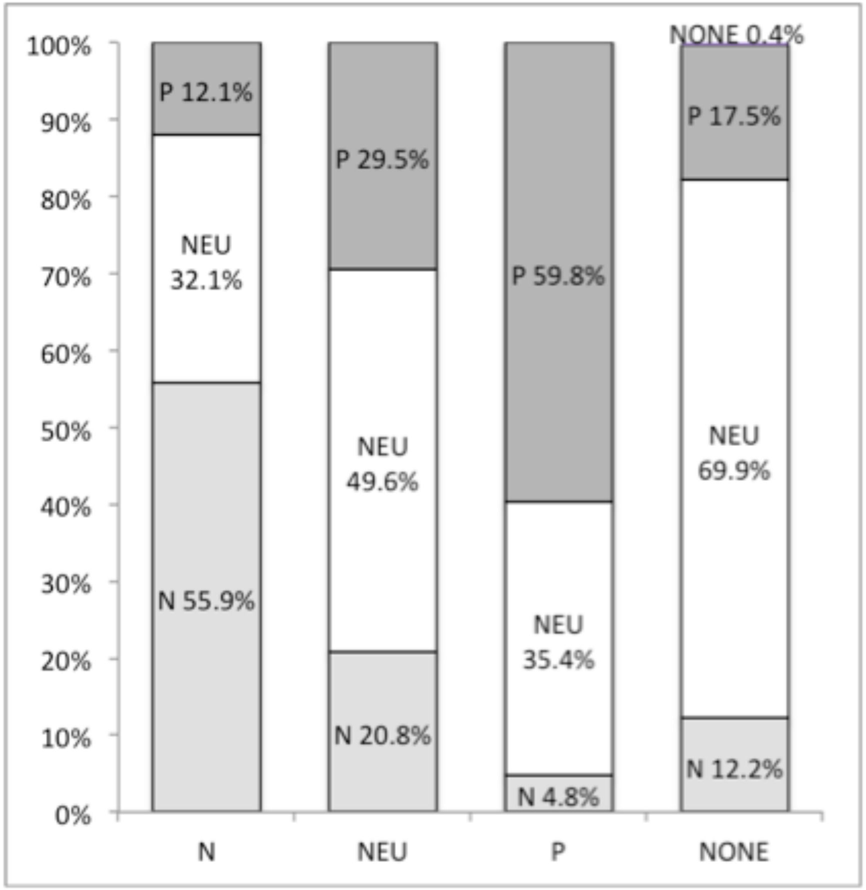
$$GSV = \frac{\left(\sum_{i=1}^{5} 2.5 \cdot i \cdot N_i + \sum_{i=1}^{5} 2.5 \cdot i \cdot P_i\right) \cdot uB}{5 \cdot (LS - NS)} \quad (1)$$

- N_i number of negative valences, placing more weight increasingly on higher intensity values.
- P_i number of positive values, weighting as above.
- *uB*: Affect Intensity.
- *LS* number of lexical segments
- *NS* number of neutral ones.

Established ranges for classification

Category	GSV Range
P+	GSV>8
P	GSV<=8
NEU	GSV<=5.4
N	GSV<=4
N+	GSV<=2
NONE	No content

Test results (3L+N)



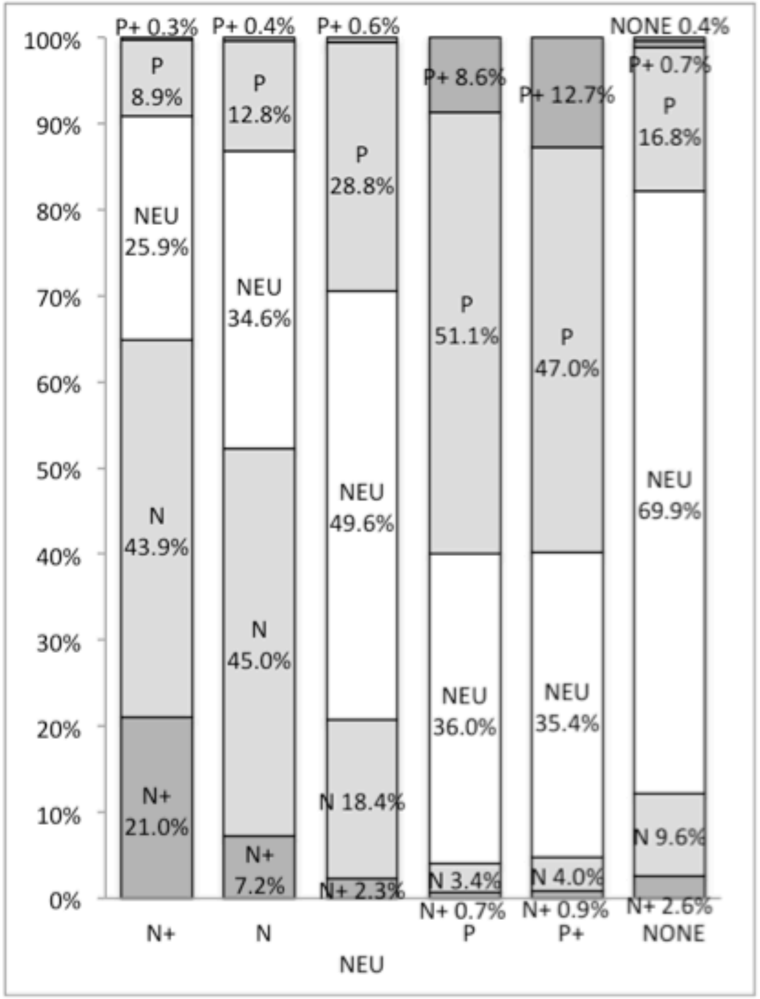
	Precision	Recall	F
N	0.559	0.691	0.618
NEU	0.497	0.023	0.043
P	0.598	0.688	0.639
NONE	0.004	1	0.008

Test results (unofficial 3L)

Actual	Pred	#	%/actual
N	*	15,840	100.00
N	N	8,848	55.86
N	NEU	5,083	32.09
N	P	1,909	12.09
NEU	*	22,713	100.00
NEU	N	2,881	12.68
NEU	NEU	15,709	69.16
NEU	P	4,123	18.15
P	*	22,231	100.00
P	N	1,068	4.80
P	NEU	7,879	35.44
P	P	13,284	59.75

	Precision	Recall	F
N	0.559	0.691	0.618
NEU	0.692	0.548	0.611
P	0.598	0.688	0.639

Test results (5L+N)



	Precision	Recall	F
N+	0.210	0.373	0.269
N	0.450	0.496	0.472
NEU	0.498	0.023	0.043
P	0.512	0.047	0.085
P+	0.127	0.885	0.223
NONE	0.004	1.000	0.008

Test results (unofficial 5L)

	N	Hits	H %	Misses	M %
N+	4,552	955	20.98	3,597	79.02
N	11,281	5,075	44.99	6,206	55.01
NEU	22,709	15,709	69.18	7,000	30.82
P	1,483	760	51.25	723	48.75
P+	20,741	2,643	12.74	18,098	87.26
Total	60,766	25,142	41.38	35,624	58.62

	Precision	Recall	F
N+	0.210	0.373	0.269
N	0.450	0.496	0.472
NEU	0.692	0.548	0.612
P	0.512	0.047	0.085
P+	0.127	0.885	0.223

Conclusions

- Very useful test for us! It confirms our suspicions that GSV is strongly influenced by length of texts (i.e., number of lexical items available). It also confirms Sentitext's middle-of-the-scale tendency.
- Twitter users employ highly emotional language, confirmed by an exceptionally high Affect Intensity average.
- The distinction between the "neutral" and "none" categories is far from clear, plus there is no practical advantage to such distinction.

Thank you!