Neutral Score Detection in Lexicon-based Sentiment Analysis: the Quartile-based Approach

Marco Vassallo[1,†], Giuliano Gabrieli[1], Valerio Basile[2] and Cristina Bosco[2]

[1]CREA Research Centre for Agricultural Policies and Bio-economy, Rome (Italy)
[2]Dipartimento di Informatica - University of Turin, Turin (Italy)

Abstract
The neutrality detection in Sentiment Analysis (SA) still constitutes an unsolved and debated issue. This work proposes an empirical method based on the quartiles of the polarity distribution for a lexicon-based SA approach. Our experiments are based on the Italian linguistic resource MAL (Morphologically-inflected Affective Lexicon) and applied to two annotated corpora. The findings provided a better detection of the neutral expressions with preserving a substantial overall polarity prediction.

Keywords
Sentiment Analysis, Lexicon, Neutrality, Optimization

1. Introduction and rationale

Sentiment Analysis (SA) is a well-studied task of Natural Language Processing (NLP), whose main objective is to classify opinions from natural language expressions as positive, neutral, negative or a mixture of those [1]. The neutrality detection in SA is an issue approached in different ways [2, 3, 4], but low agreement on how detecting neutral expressions still exists [4, p.136]. In this paper, we approach neutrality detection in lexicon-based SA, where an affective lexicon provides polarity scores ranging from $-a$ to $+a$ with $a \in N$, by using a descriptive statistical method based on the quartiles.

To our knowledge, this issue was not investigated so far. We aim at drawing attention towards a better prediction of the neutral expressions. This is done by automatically finding out an optimal interval of neutral scores with a control for the asymmetry of the distribution of the scores across the polarity spectrum. Traditionally, neutrality scores have been assumed to be around point 0, or within a conventionally fixed and algebraically-led interval of $[-.5; +.5]$. Conversely, it seems more reasonable to postulate that this neutral cluster should lie in a dynamic interval around the zero value. As expected, the $[-.5; +.5]$ interval is indeed insufficient for capturing the neutral values, especially when the polarity scores are symmetrical around the point zero. This is because small positive or negative deviations from zero can be incorrectly classified into their respective polarity if they are neutral. Furthermore, for topics with many controversial opinions, where polarizaties are indeed dispersed, the misclassification of neutral expressions appears significant, as small positive and negative deviations from zero might be more frequent. As a consequence, the neutral interval also appears to be topic-oriented and thus differs from any SA task, as the topic could, in turn, also influence the symmetry of the distribution of scores. The linguistic counterpart to this phenomenon is that "opinions may be so different that common ground may not be found" [5].

On the other hand, especially in the case of unimodal distributions, the more asymmetrical the polarity scores distribution is, the more the polarities might be positively or negatively skewed, and the less likely a false neutral classification should occur. In the case of multimodal distributions, with multiple possible polarizations, detecting the asymmetry becomes more complex as well as the neutral expressions. But, despite the peculiar situation with the same frequencies for oppositely polarized scores, the more a multimodal distribution is skewed (many different modes/peaks possibly far from zero) the less likely false neutral classifications should again occur.

2. The quartile-based approach

The quartiles are the values of a variable that divide its relative distribution into four equal parts once the data are arranged in ascending order. These values are as follows: the first quartile $Q1$ represents the value below which 25% of the data are situated; $Q2$ is the second quartile or the Median value that exactly splits the data into two halves; $Q3$, the third quartile, is the value above which 25% of the data is situated.
Considering that lexicon-based SA provides a range of

CLiC-it 2024: Tenth Italian Conference on Computational Linguistics, Dec 04 — 06, 2024, Pisa, Italy
*Corresponding author.
marco.vassallo@crea.gov.it (M. Vassallo); giuliano.gabrieli@crea.gov.it (G. Gabrieli); valerio.basile@unito.it (V. Basile); cristina.bosco@unito.it (C. Bosco)
0000-0001-7016-6549 (M. Vassallo); 0000-0001-8110-6832 (V. Basile); 0000-0002-8857-4484 (C. Bosco)

scores from $-a$ to $+a$ (with $a \geq 1$) the neutral scores should reasonably fall into a sub-interval that belongs to $[Q1; Q3]$ and possibly includes the absolute zero (the neutral score by intuition). Furthermore, this sub-interval of neutral scores is, reasonably, sensitive to the topic and therefore to the asymmetry of the entire polarity distribution. Quartiles also take into account the potential asymmetry of a data distribution since typical values of skewed data fall between $Q1$ and $Q3$. To understand this asymmetrical process, and thus the usefulness of the quartiles in detecting potential deviation from symmetry in a data set, we recall the Galton Skewness index, also known as Bowley's skewness index [6], that is based on the quartiles and defined as follows:

$$G = [(Q3 - Q2) - (Q2 - Q1)]/(Q3 - Q1)$$

G measures the level of skewness in the dataset as the difference between the lengths of the upper quartile ($Q3 - Q2$) and the lower quartile ($Q2 - Q1$), normalized by the length of the interquartile range ($Q3 - Q1$), i.e. a measure of the variability of the data from the median ($Q2$). The G index ranges from -1 (the distribution is negatively skewed) to +1 (the distribution is positively skewed) and it is zero for a symmetric distribution.

The logic of the optimal quartile-based interval The main challenge now is to reveal the sub-interval skewed-variant within $[Q1; Q3]$ that can predict the true neutral scores without decreasing the positive and negative predictions. By searching for true neutral scores, at the same time we risk increasing false positives and negatives. This is what presumably happens whenever a default neutral interval of $[-.5; +.5]$ is selected. The computational idea is straightforward and intuitive, and it makes use of annotated corpora. Once calculating the $Q1$ and $Q3$ in the polarity scores distribution, a R-script is set up to routinize a computational process starting from the interval $[0; 0]$ to $[Q1; Q3]$ in increasing/decreasing steps of .005 for stopping to a sub-interval (within $[Q1; Q3]$) that simultaneously optimized the F1 score for the neutral, positive and negative classes. If this simultaneous optimization yields to acceptable F1-scores the entire proposed process can be considered sufficient. In order to validate the approach and provide a tool that can be applied to unseen data, we implemented a cross-validation experiment. We randomly split each dataset into training and test sets by varying percentages of both in steps of 10%. The strategy of the dual portion-variant steps was due to the rationale of considering all potential and reasonable unseen data situations. The logic steps of the optimal quartiles-based interval was then run on every split to find those optimal intervals in conformity with those desiderata percentages of training and test. It is straightforward to notice that the optimal intervals of the cross-validation might not coincide with those found in the whole initial dataset. Nevertheless, they can provide a validation range to which the initial optimal intervals are the upper bound.

3. Experiments on two corpora

We considered two datasets:

- AGRITREND [7], a corpus of Italian tweets on general agricultural topics manually annotated by three different annotators
- SENTIPOLC which is the benchmark dataset used in the SENTIment Polarity Classification shared task held in EVALITA 2016 [8], a challenge on polarity detection on Italian tweets; this is another annotated corpus of Italian tweets including texts for three different topics (i.e., general (GEN), political (POL) and sociopolitical (SPOL)).

The SENTIPOLC dataset is composed of 9,410 tweets, pre-divided into a training set (7,410 tweets) and a test set (2,000 tweets). The annotation scheme of SENTIPOLC comprises two non-mutually exclusive binary labels for positive and negative polarity, It is therefore possible for a tweet to be marked as neutral (non-positive and non-negative) or mixed (positive and negative at the same time). Other two binary labels mark the subjectivity of the message (subjective vs. objective) and the ironic content. Finally, an additional layer of annotation labels the literal positivity and negativity of the tweet, which could be different from the actual polarity (called "overall" polarity in SENTIPOLC). Note that, while this scheme is quite flexible, not all possible combinations of labels are allowed. In particular, according to a rule for the dataset, a tweet cannot be labeled at the same time as objective and as displaying sentiment polarity or irony. The origin of the tweets in SENTIPOLC is diverse, with 6,421 tweets which were part of the corpus collected for the previous edition of the shared task [9], and the rest from other smaller collections or drawn from Twitter especially for the purpose of organizing SENTIPOLC 2016. The annotation scheme of AGRITREND is exactly the same as SENTIPOLC by design.

For this experiment, we applied the MAL[1] (Morphologically-inflected-Affective-Lexicon) [7] as affective lexicon ranging from -1 to 1. It was originally

[1]The MAL was also further implemented with a weighted version named W-MAL [10] ranging from -5.16 to 5.95 that has considered the word frequencies of TWITA [11]. We also applied W-MAL in this experiment and the results were in line with those of MAL, although even more extreme. However, since the W-MAL was updated until 2020 and the datasets of AGRITREND and SENTIPOLC were respectively collected until 2022 and 2016, we prefer to present results from the unweighted version.

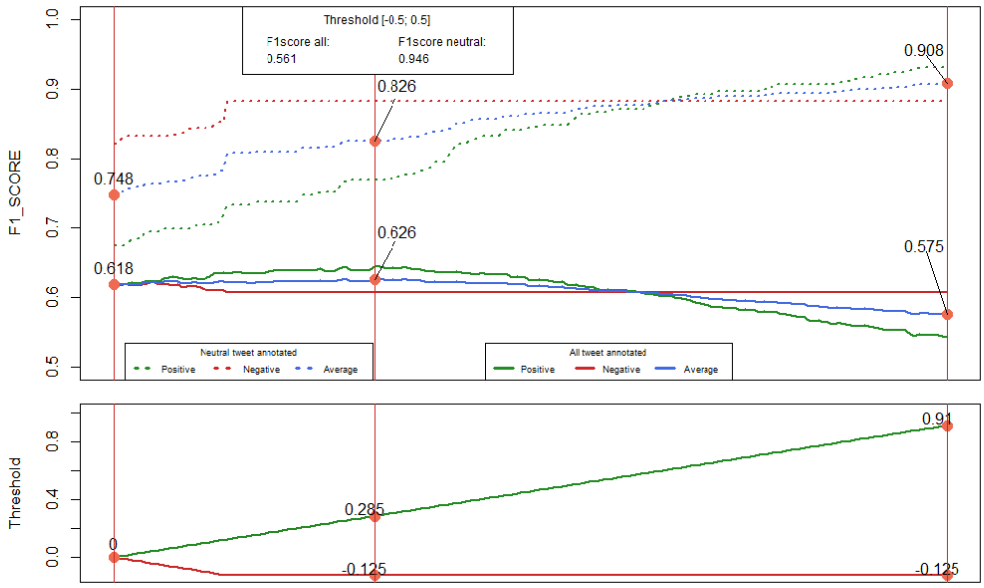

Figure 1: Results of the polarity classification on AGRITREND - F1 scores

derived from Sentix [12] and successively augmented with a collection of Italian forms from the Morph-It [13]. Since the MAL does not classify the mixed labels, we selected the tweets with positive, negative and neutral polarities from both datasets. As a result, AGRITREND was finally composed of 1,224 tweets with 171 neutral annotated expressions, while SENTIPOLC of 8,892 tweets with 3713 neutral annotated expressions also topic-classified as follows: 1,537 for the GEN topic; 1,510 for the POL topic; 666 for SPOL topic.

3.1. Results on AGRITREND

Corpus	Q1	Q2	Q3	G
AGRITREND	-0.125	0.280	0.907	0.215
SENTIPOLC ALL	0.099	0.656	1.315	0.084
SENTIPOLC GEN	0.000	0.533	1.160	0.081
SENTIPOLC POL	0.269	0.816	1.470	0.090
SENTIPOLC SPOL	0.060	0.589	1.193	0.066

Table 1
Quartiles and G values

In Table 1, the quartiles and G values are reported. It can be observed that AGRITREND scores are slightly skewed positively (i.e., the G is 0.215).
Figure 1 shows the computational optimization of the quartile-based approach. Starting from the right side of the figure, this corpus has $[Q1; Q3] = [-0.125; 0.907]$ that corresponds to an average F1 score of 0.908 for neutral and 0.575 for positive/negative with negative higher than positive. Setting the threshold for neutral to the default values of $[-0.5; 0.5]$ (i.e., in correspondence of the box on top of the figure) the F1 score (on average) for neutral increases to 0.946, but the F1 score (on average) for positive/negative decreases to 0.561. Similarly, at the zero point, F1-scores are on average 0.618 and 0.748. By triggering the optimization process from $[0; 0]$, it converges to the optimal interval of $[-0.125; 0.285]$, where F1 scores (on average) are 0.826 for neutral and 0.626 for positive/negative. This result represents a better trade-off for a simultaneous prediction of all the labels with respect to using the default or the zero point intervals.

Tables 2–6 report the quartile-based approach (Table 2 for AGRITREND) cross-validation results with training and test set steps strategy. The optimal interval initially found of $[-0.125; 0.285]$ can be confirmed from 90%-10% to 80%-20% step of training and test sets percentages split. However, it would be possible to move until 60%-40% split level (highlighted in bold) which was the optimal interval range that simultaneously optimized the F1 score for the neutral, positive and negative classes across the cross-validation. In this case, the upper limits increase and thus they need to be looked into. The F1-scores (on average) for the training set range from 0.626 to 0.630 and from 0.827 to 0.849 for polarized and neutral scores, respectively. The F1-scores (on average) for the test set range from 0.624 to 0.628 and from 0.827 to 0.829 for polarized and neutral scores, respectively. Table 9 presents examples of polarized tweets annotated

% Train	% Test	Training				Test			
		Limit		F1-score		Limit		F1-score	
		Lower	Upper	Avg. all	Avg. Neutral	Lower	Upper	Avg. all	Avg. Neutral
10	90	-0,250	0,320	0,6157	0,8736	-0,075	0,125	0,6170	0,8435
20	80	-0,135	0,225	0,6358	0,8421	-0,035	0,035	0,6226	0,7856
30	70	-0,160	0,225	0,6368	0,8218	-0,070	0,070	0,6304	0,7758
40	60	-0,140	0,250	0,6303	0,8255	-0,135	0,160	0,6337	0,8127
50	50	-0,130	0,250	0,6286	0,8287	-0,070	0,070	0,6255	0,7768
60	**40**	**-0,125**	**0,320**	**0,6258**	**0,8492**	**-0,125**	**0,305**	**0,6243**	**0,8293**
70	30	-0,125	0,320	0,6284	0,8375	-0,125	0,285	0,6221	0,8247
80	20	-0,125	0,285	0,6297	0,8259	-0,125	0,285	0,6237	0,8191
90	10	-0,125	0,285	0,6299	0,8269	-0,125	0,315	0,6285	0,8266

Table 2
Training and test sets - Optimal quartile-based intervals - AGRITREND

% Train	% Test	Training				Test			
		Limit		F1-score		Limit		F1-score	
		Lower	Upper	Avg. all	Avg. Neutral	Lower	Upper	Avg. all	Avg. Neutral
10	90	0	1,295	0,5535	0,8812	0	1,200	0,5679	0,8820
20	80	0	1,295	0,5568	0,8926	0	1,075	0,5470	0,8648
30	70	0	1,310	0,5558	0,8929	0	1,165	0,5445	0,8700
40	60	0	1,320	0,5584	0,8913	0	1,165	0,5411	0,8693
50	50	0	1,320	0,5559	0,8874	0	1,165	0,5435	0,8670
60	40	0	1,310	0,5554	0,8853	0	1,165	0,5439	0,8661
70	**30**	**0**	**1,210**	**0,5516**	**0,8740**	**0**	**1,165**	**0,5474**	**0,8673**
80	20	0	1,175	0,5501	0,8700	0	1,165	0,5478	0,8683
90	10	0	1,165	0,5472	0,8685	0	1,165	0,5489	0,8699

Table 3
Training and test sets - Optimal quartile-based intervals - SENTIPOLC - ALL

% Train	% Test	Training				Test			
		Limit		F1-score		Limit		F1-score	
		Lower	Upper	Avg. all	Avg. Neutral	Lower	Upper	Avg. all	Avg. Neutral
10	90	0	0,535	0,5572	0,7956	0	0,500	0,5711	0,7830
20	80	0	0,535	0,5807	0,8072	0	1,100	0,5573	0,8510
30	70	0	0,520	0,5747	0,7937	0	0,450	0,5615	0,7651
40	60	0	0,520	0,5809	0,7941	0	1,175	0,5658	0,8662
50	50	0	0,530	0,5774	0,7903	0	0,770	0,5693	0,8275
60	40	0	0,530	0,5764	0,7897	0	1,085	0,5695	0,8598
70	30	0	1,010	0,5768	0,8594	0	1,085	0,5707	0,8591
80	**20**	**0**	**0,520**	**0,5747**	**0,7850**	**0**	**1,085**	**0,5693**	**0,8593**
90	10	0	1,010	0,5722	0,8545	0	1,085	0,5737	0,8627

Table 4
Training and test sets - Optimal quartile-based intervals - SENTIPOLC - GEN

% Train	% Test	Training				Test			
		Limit		F1-score		Limit		F1-score	
		Lower	Upper	Avg. all	Avg. Neutral	Lower	Upper	Avg. all	Avg. Neutral
10	90	0	1,370	0,5395	0,8897	0	1,440	0,5322	0,8872
20	80	0	1,430	0,5531	0,8957	0	1,410	0,5267	0,8835
30	**70**	**0**	**1,440**	**0,5537**	**0,8945**	**0**	**1,300**	**0,5203**	**0,8724**
40	60	0	1,440	0,5582	0,8949	0	1,410	0,5147	0,8904
50	50	0	1,440	0,5553	0,8960	0	1,410	0,5210	0,8918
60	40	0	1,440	0,5529	0,8965	0	1,410	0,5248	0,8928
70	30	0	1,440	0,5458	0,8992	0	1,350	0,5309	0,8843
80	20	0	1,440	0,5404	0,8971	0	1,445	0,5338	0,8950
90	10	0	1,440	0,5385	0,8960	0	1,445	0,5367	0,8951

Table 5
Training and test sets - Optimal quartile-based intervals - SENTIPOLC - POL

% Train	% Test	Training				Test			
		Limit		F1-score		Limit		F1-score	
		Lower	Upper	Avg. all	Avg. Neutral	Lower	Upper	Avg. all	Avg. Neutral
10	90	-0,025	1,470	0,5277	0,8947	0,000	1,315	0,5969	0,8976
20	80	0,000	1,255	0,5229	0,8758	0,000	1,280	0,5921	0,8971
30	70	0,000	1,215	0,5146	0,8824	0,000	1,195	0,5818	0,8916
40	60	0,000	1,215	0,5186	0,8821	0,000	1,185	0,5760	0,8931
50	50	0,000	1,210	0,5247	0,8763	0,000	1,185	0,5732	0,8942
60	40	0,000	1,205	0,5306	0,8799	0,000	1,165	0,5671	0,8865
70	**30**	**0,000**	**1,190**	**0,5331**	**0,8812**	**0,000**	**1,180**	**0,5634**	**0,8864**
80	20	0,000	1,165	0,5377	0,8828	0,000	1,180	0,5551	0,8863
90	10	0,000	1,165	0,5436	0,8828	0,000	1,170	0,5520	0,8826

Table 6
Training and test sets - Optimal quartile-based intervals - SENTIPOLC - SPOL

as neutral and correctly classified by the quartile-based approach.

3.2. Results on SENTIPOLC

Domains	low	up	F1-AVG	F1-Neutral
GEN	0	0.52	0.570	0.784
POL	0	1.44	0.538	0.895
SPOL	0	1.19	0.548	0.884

Table 7
The optimal quartile-based intervals and F1-scores in SENTIPOLC domains

Domain	AVG-[-.5;.5]	Neutral-[-.5;.5]	AVG-zero	Neutral-zero
GEN	0.567	0.923	0.520	0.651
POL	0.507	0.925	0.403	0.605
SPOL	0.507	0.923	0.432	0.614

Table 8
F1-scores for the zero and [-.5 +.5] intervals in SENTIPOLC domains

The values in Table 1 show that the polarized score distribution is quite symmetrical even within each domain (i.e., the G values are all close to 0). The results on SENTIPOLC All (i.e., with no specific domain) showed an optimal interval of $[0; 1.175]$ with 0.548 and 0.868 of F1-score (on average) for positive/negative and neutral, respectively. In comparison to the default values of the interval $[-0.5; 0.5]$ and to the zero point, the F1-score (on average) for positive/negative also increases here (from 0.526 and 0.455 to 0.549) while preserving a high F1-score of 0.870 for the neutrals. When the polarized scores distribution is close to perfect symmetry, the difference between $[Q1; Q3]$ and the optimal interval is minimal, which is expected because the quartiles are skew-dependent.

When the SENTIPOLC dataset is divided in specific domains, the optimal quartile-based intervals confirmed the best balance of the predictions between positive/negative and neutral scores across all domains (see F1-scores in Table 7 vs Table 8). Interestingly, the effect of the optimization process is more visible on the specific topics POL and SPOL of SENTIPOLC (Tables 5 and 6) across the cross-validation process. Even better for POL domain where at least 30/% of training would be necessary (Table 5). This could be due to the topic being more specific with a higher likelihood of finding neutral expressions. As shown also in Tables 7 and 8, the F1-scores for the neutral expressions are higher both for POL and SPOL than those of GEN. Concerning this latter, the results in table 4 indicate a kind of over-fitting. This may make sense, considering that this section of the dataset, being open-domain, has likely a higher degree of lexical variation. Furthermore, the recall index was even found higher for the test set than the one of the training set.

4. Discussion

In this work, we proposed a descriptive statistical method for a better detection of the neutral expressions in lexicon-based SA with polarity scores. This method is based on quartiles and therefore on the assumption that an optimal interval for neutral scores should take always into account the potential asymmetry of the polarity distribution. This seems also in line with the linguistic speculation that the less a topic looks polarized the more difficult it should be to detect neutral expressions. The rationale is that even small positive or negative values around the zero point could be classified as such while they should be instead neutral. Conversely, the more a topic looks polarized, the easier it should be to detect neutral expressions. In our view, an optimal interval for detecting neutral scores in lexicon-based SA should control for biases caused by the symmetry unbalance in polarity predictions.

The optimization process we presented starts with computing the first ($Q1$) and the third ($Q3$) quartiles of a polarity score distribution and afterwards finding out the optimal interval within $[Q1, Q3]$ that balances the polarity and the neutral predictions simultaneously. We

Original text	Bag of words	MAL score
A. #Grow!2019: i produttori agricoli #Agrinsieme si confrontano sul #trasporto su gomma e portuale; interventi del copresidente del coordinamento @dinoscanavino e dell'Ad di #Acea	produttori agricoli confrontano gomma portuale interventi copresidente coordinamento	-0.0061
A. Ortofrutta, analisi dei consumi durante il coronavirus-Uci-Unione Coltivatori Italiani https://t.co/UKOaone6oJ	analisi consumi coronavirus unione coltivatori italiani	0.201
S. Italia progredisce se parla di innovazione, scuola digitale e alternanza scuola-lavoro #labuonascuola @cittascienza http://t.co/2pR7MVw40F	Italia progredisce parla innovazione scuola digitale alernananza scuola lavoro	0.229
S. Come la tecnologia può cambiare le scuole e il sistema di apprendimento? #scuola #labuonascuola http://t.co/9bD4YsA2aG	tecnologia cambiare scuole sistema apprendimento	0.423

Table 9
Examples of polarized tweets from *AGRITREND **A.** and SENTIPOLC **S.*** correctly detected as neutral by the quartile-based approach.

demonstrated that when the topic of a corpus is generic it requires at least 60%-70% of the data as the training set to find out the optimal interval of neutrals. On the other hand, the more specific the topic is, the less training data it requires to achieve a reasonable optimal interval for neutrals. We stipulate that even a 30% split might be sufficient. Our results on two datasets are promising in providing a more precise prediction of neutral scores while preserving a good polarity prediction in comparison to the one obtained by the usual interval of $[-.05; +.05]$ and by the single zero point.

5. Conclusion and future work

The asymmetry of a polarity scores distribution seems to be topic-oriented and therefore the neutrality detection for a lexicon-based SA with polarity scores reasonably passes through an optimal interval within the first and the third quartile $[Q1, Q3]$ that takes this asymmetry into account. The findings of this work stipulated that the quartile-based approach is suitable for any corpus where a task of lexicon-based SA with scores is performed. Hence, we do strongly recommend further experiments on other corpora, both annotated and unannotated, and comparing/integrating this method with others (e.g. Valdivia et al. [4]) for the common objective of detecting neutral expressions. Eventually, it is worthwhile noticing that our methodological framework led us to run experiments on test sets of different sizes in order to consider all potential and reasonable unseen data situations. Alternatively, one could propose a similar experiment with fixed-size test sets, which would have provided more stable, comparable results even with established benchmarks, but on the other hand would also significantly reduce the amount of test data

References

[1] S. Sun, C. Luo, J. Chen, A review of natural language processing techniques for opinion mining systems, Information Fusion 36 (2017) 10–25. URL: https://www.sciencedirect.com/science/article/pii/S1566253516301117. doi:https://doi.org/10.1016/j.inffus.2016.10.004.

[2] M. Koppel, J. Schler, The importance of neutral examples for learning sentiment., Computational Intelligence 22 (2006) 100–109. doi:10.1111/j.1467-8640.2006.00276.x.

[3] B. Pang, L. Lee, Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales, in: K. Knight, H. T. Ng, K. Oflazer (Eds.), Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05), Association for Computational Linguistics, Ann Arbor, Michigan, 2005, pp. 115–124. URL: https://aclanthology.org/P05-1015. doi:10.3115/1219840.1219855.

[4] A. Valdivia, M. V. Luzón, E. Cambria, F. Herrera, Consensus vote models for detecting and filtering neutrality in sentiment analysis, Information Fusion 44 (2018) 126–135. URL: https://www.sciencedirect.com/science/article/pii/S1566253517306590. doi:https://doi.org/10.1016/j.inffus.2018.03.007.

[5] N. Koudenburg, Y. Kashima, A polarized discourse: Effects of opinion differentiation and structural differentiation on communication, Personality and Social Psychology Bulletin 48 (2022) 1068–1086. URL: https://doi.org/10.1177/01461672211030816. doi:10.1177/01461672211030816, pMID: 34292094.

[6] A. Bowley, Elements of Statistics, Studies in economics and political science, P. S. King & son, 1917. URL: https://books.google.it/books?id=M4ZDAAAAIAAJ.

[7] M. Vassallo, G. Gabrieli, V. Basile, C. Bosco, The tenuousness of lemmatization in lexicon-based sen-

timent analysis, in: Proceedings of the Sixth Italian Conference on Computational Linguistics - CLiC-it 2019, Academia University Press, 2019.

[8] F. Barbieri, V. Basile, D. Croce, M. Nissim, N. Novielli, V. Patti, Overview of the Evalita 2016 SENTIment POLarity Classification Task, in: Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016), CEUR-WS.org, 2016.

[9] V. Basile, A. Bolioli, M. Nissim, V. Patti, P. Rosso, Overview of the Evalita 2014 SENTIment POLarity Classification Task, in: Proceedings of the 4th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'14), Pisa, Italy, 2014. URL: https://inria.hal.science/hal-01228925. doi:10.12871/clicit201429.

[10] M. Vassallo, G. Gabrieli, V. Basile, C. Bosco, Polarity imbalance in lexicon-based sentiment analysis, in: Proceedings of the Seventh Italian Conference on Computational Linguistics - CLiC-it 2020, 2020, pp. 457–463. doi:10.4000/books.aaccademia.8964.

[11] V. Basile, M. Lai, M. Sanguinetti, Long-term Social Media Data Collection at the University of Turin, in: Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), CEUR-WS.org, 2018.

[12] V. Basile, M. Nissim, Sentiment analysis on Italian tweets, in: Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, 2013, pp. 100–107.

[13] E. Zanchetta, M. Baroni, Morph-it! A free corpus-based morphological resource for the Italian language, in: Proceedings of Corpus Linguistics 2005, 2006.