Improving Sentence-level Subjectivity Classification through Readability Measurement

Robert Remus
Dept. of Computer Science, NLP Group
University of Leipzig, Germany
rremus@informatik.uni-leipzig.de

Abstract

We show that the quality of sentence-level subjectivity classification, i.e. the task of deciding whether a sentence is subjective or objective, can be improved by incorporating hitherto unused features: readability measures. Hence we investigate in 6 different readability formulae and propose an own. Their performance is evaluated in a 10-fold cross validation setting using machine learning. Thereby, it is demonstrated that sentence-level subjectivity classification benefits from employing readability measures as features in addition to already well-known subjectivity clues.

1 Introduction

Wiebe et al. (2004) refer to *subjectivity* in natural language as "aspects of language used to express opinions, evaluations, and speculations". For example, an utterance like

> "In the end, though, it is only mildly amusing when it could have been so much more."

clearly bears an opinion, i.e. is subjective, whereas an utterance like

> "The movie takes place in mexico, 2002."

clearly does not[1]. *Readability* is usually refered to as "the degree to which a given class of people find certain reading matter compelling and, necessarily, comprehensible" (cf. McLaughlin (1969)). So whereas the meaning of a sentence like

> "Nanometer-sized single crystals, or single-domain ultrafine particles, are often referred to as nanocrystals."

is quite difficult to grasp, a sentence like

> "Wills and Kate get into marriage mode."

is much easier to understand[2]. Why is that? The former sentence not only exhibits a more complex syntactic structure than the latter, but also extensively utilises domain-specific terminology that many readers would not be familiar with.

Although counter-intuitive on first sight, we pose the following hypothesis: There is a *connection between subjectivity and readability* in natural language text. If so, we may assume that knowing about its readability possibly yields valuable information regarding its subjectivity.

1.1 Related Work

To our best knowledge, readability measures have not been used to assess the subjectivity of any lexical units so far, be it word forms, phrases, sentences or whole documents. However, there is Hoang et al. (2008)'s work on evaluating the *quality* of user-created documents, and recent work on grading the *helpfulness* of reviews by (O'Mahony and Smyth, 2010), both incorporating readability measures. Close to our research is Nishikawa et al. (2010)'s study on *sentiment summarisation* which utilises measures both for informativeness and readability. Very recent support in favour of our hypothesis is provided by (Lahiri et al., 2011), who measure a correlation between *informality* and readability.

As subjectivity classification poses many interesting challenges and has many applications in NLP including genre detection, flame recognition and information extraction, the identification of

[1] Both sentences are taken from (Pang and Lee, 2004)'s *subjectivity data set v1.0.*

[2] The first sentence is taken from the Wikipedia article http://en.wikipedia.org/Nanoparticles, the second is taken from Sun Daily News' homepage http://www.thesun.co.uk, both accessed on January 8th, 2011.

Bolette Sandford Pedersen, Gunta Nešpore and Inguna Skadiņa (Eds.)
NODALIDA 2011 Conference Proceedings, pp. 168–174

features for subjectivity classification and the classification itself has been extensively studied. Especially Wiebe et al. contributed a lot to the field: Wiebe (2000) learns subjective adjectives from corpora in a semi-supervised fashion, while Wiebe et al. (2001) identify other subjectivity clues using collocations. Riloff et al. (2003) and Riloff and Wiebe (2003) present ways to mine extraction patterns for subjective expressions. Wiebe et al. (2004) summarise these findings and show how different features work "together in concert". Wiebe et al. (2005) introduce a scheme for annotating opinions and the like in Wiebe et al. (2003)'s English-language *Multi-Perspective Question Answering* (MPQA) corpus. Wilson et al. (2004) assess the strength of deeply-nested opinions. Wiebe and Riloff (2005) create high-precision classifiers for distinguishing between subjective and objective sentences and use them as a source for learning additional subjectivity clues.

Yu and Hatzivassiloglou (2003) perform both document- and sentence-level subjectivity classification using Naïve Bayes classifiers and several unsupervised approaches. Pang and Lee (2004) use a graph-based formalism to first tell subjective and objective sentences apart, then perform a polarity classification employing both Naïve Bayes classifiers and Support Vector Machines.

1.2 Outline

This paper is structured as follows: In the next section, we describe our method by presenting hitherto unused features for subjectivity classification: readability measures. In Section 3 we describe our experimental setup and evaluate its performance. Finally, we draw conclusions and point out possible directions for future work in Section 4.

2 Method

Following our assumption that knowing about the readability of natural language text possibly yields valuable information regarding its subjectivity, we will try to measure readability and later exploit this information for *sentence-level subjectivity classification*.

According to Klare (1974)'s survey there are 3 possible solutions to "tell whether a particular piece of writing is likely to be readable to a particular group of readers": A first solution is simply to guess. A second solution are tests, manually built and refined. A third solution are *readability measures*. We will assess readability by such measures for the obvious reason that some of them are automatically computable.

2.1 Readability Measures

We chose 6 different readability formulae from the large body of available readability measures. All measures we chose are automatically computable and *do not* depend on lexical resources like word lists and the like. We solely present the 6 formulae themselves. The reader interested in their underlying ideas, their development and the derivation of their constants and variables may be refered to the aforementioned Klare (1974), or the original work cited below.

Additionally, we propose an easy to calculate formula that embodies our own intuition for assessing readability.

2.1.1 Devereux Readability Index

The *Devereux Readability Index* D was introduced by Smith (1961) and is calculated as shown in Equation 1,

$$\mathrm{D} = 1.56 \times wl + 0.19 \times sl - 6.49 \qquad (1)$$

where wl is the average word length in characters and sl is the average sentence length in words. The Devereux formula was designed to cover school grades 4 to 12. Hence, the higher the value of D, the less readable the graded text according to the formula.

2.1.2 Easy Listening Formula

The *Easy Listening Formula* EL was introduced by Fang (1966) and is calculated simply as shown in Equation 2,

$$\mathrm{EL} = npsw \qquad (2)$$

where $npsw$ is the average number of polysyllabic words per sentence, i.e. words with more than just one syllable. The Easy Listening Formula is, as the name suggests, tailored to "listenability" rather than readability. Therefore, the higher the value of EL, the less "listenable" the graded text according to the formula.

2.1.3 Fog Index

The *Fog Index* was introduced in Gunning (1952) and reformulated by Powers et al. (1958). It is calculated as shown in Equation 3,

$$\mathrm{FI} = 3.068 + 0.0877 \times sl + 0.0984 \times nosw \quad (3)$$

where sl is the average sentence length in words and $nosw$ is the average number of one-syllable words per sentence. The higher the value of FI, the less readable the graded text according to the formula.

2.1.4 FORCAST

The *FORCAST* formula F was introduced in Caylor et al. (1973). It is calculated as shown in Equation 4,

$$\text{F} = 20.41 - 0.11 \times nosw \tag{4}$$

where $nosw$ is the average number of one-syllable words per sentence. The higher the value of F, the less readable the graded text according to the formula.

2.1.5 New Reading Ease Index

The *New Reading Ease Index* NREI was introduced by Farr et al. (1951) and is calculated as shown in Equation 5,

$$\text{NREI} = 1.599 \times nosw - 1.015 \times sl - 31.517 \tag{5}$$

where $nosw$ is the average number of one-syllable words per sentence and sl is the average sentence length in words. The higher the value of NREI, the less readable the graded text according to the formula.

2.1.6 SMOG

The *SMOG* grading S was introduced by McLaughlin (1969) and is calculated as shown in Equation 6,

$$\text{S} = 3 + \sqrt{npsw} \tag{6}$$

where $npsw$ is the number of polysyllabic words. Again, the higher the value of SMOG, the less readable the graded text according to the formula.

2.1.7 An Own Formula

Following our intuition for how to assess readability, we propose an easy to calculate formula, shown in Equation 7.

$$\text{W} = wl \times sl \times ntop \tag{7}$$

Here, wl is the average word length per sentence, sl is the average sentence length in words and $ntop$ is the average number of words in each sentence, that are not among the top 1,000 most frequent words of a large reference corpus. This list was automatically extracted from an English-language newspaper corpus of University of Leipzig's *Wortschatz*[3] project consisting of 49,628,893 distinct sentences, 4,785,862 word types and 926,766,504 word tokens. The idea behind $ntop$ is, that high-frequency words are common to all readers, whereas medium- to low-frequency words are not necessarily. The more unfamiliar words a reader encounters in a text, the less readable it is. wl and sl basically capture the same idea: both longer sentences and longer words lead to less readable text. Just as for all the other formulae, the higher the value of W, the less readable the graded text according to our formula.

	D	EL	FI	F	NREI	S	W
wl	×						×
sl	×		×		×		×
$nosw$			×	×	×		
$npsw$		×				×	
$ntop$							×

Table 1: Comparison of language characteristics captured by the presented readability formulae. wl denotes the average word length, sl the average sentence length in words, $nosw$ the average number of one-syllable words per sentence, $npsw$ the average number of polysyllabic words per sentence.

2.1.8 Other Formulae

There are other well-known readability formulae we did not investigate in yet, e.g. Lorge (1939) and Lorge (1948)'s *Lorge formula*, Flesch (1944)'s *Flesch formula* and Dale and Chall (1948)'s *Dale-Chall formula*. These rely on lexical resources some of which are not publicly available and additionally introduce stronger language dependency.

2.2 Summary

Different readability formulae capture different language characteristics, as summarised in Table 1. Not only do they differ in their encoded features, but also in their intended outcome. Whereas some aim to determine a school grade, some refer to tables for further interpretation. For those reasons, the readabilities calculated by the presented measures are not comparable in general, although they do have in common, that higher values signalise less readable (or less listenable) text.

[3] http://wortschatz.informatik.uni-leipzig.de/

3 Evaluation

In order to evaluate whether the presented readability measures indeed yield possibly valuable information regarding a natural language text's subjectivity or not, we perform a sentence-level subjectivity classification using readability formulae as features: i.e., given a sentence, extract its features and classify it as being either subjective or objective.

3.1 Experimental Setup

The evaluation data set, the features and the text classifier we used in our experiments are now briefly described.

3.1.1 Evaluation Data Set

To ensure comparability and reproducibility of our results we use Pang and Lee (2004)'s publicly available *subjectivity data set v1.0*[4]. This widely-used data set consists of 5,000 sentences marked as "subjective" and 5,000 sentences marked as "objective".

3.1.2 Features

Our baseline features are *subjectivity clues* provided by Wilson et al. (2005). Their freely available English-language lexical resource encompasses 8,221 word forms, each manually annotated for being either a strong or a weak subjectivity clue and for its polarity being either positive, negative or neutral. We only used the 5,569 *strong* subjectivity clues to form a solid baseline. Examples of strong subjectivity clues include *disagree*, *love* and *overstate*.

Our additional features are the readability formulae presented in Section 2. Although most of them were developed to capture readability of whole texts, we apply them to single sentences. Their minima, maxima, averages and standard deviances measured in the 5,000 subjective and 5,000 objective sentences are shown in Table 4.

3.1.3 Text Classifier

The actual text classification is performed by Support Vector Machines (SVMs) (cf. Vapnik (1995) and Cortes and Vapnik (1995)). SVMs are known for being able to handle large feature spaces while simultaneously limiting overfitting. Because of the large number of baseline features at hand, SVMs were a natural choice. The SVMs were trained using an radial basis function kernel as provided by *LibSVM* (cf. Chang and Lin (2001)).

Feature(s)	P	R	F
Clues	**0.849**	0.158	0.263
D	0.695	0.558	0.618
EL	0.654	0.443	0.527
FI	0.652	**0.868**	**0.743**
F	0.524	0.767	0.623
NREI	0.651	0.859	0.739
S	0.603	0.810	0.680
W	0.630	0.754	0.685

Table 2: Precision P, recall R and f-score F of all single features.

Feature(s)	P	R	F
EL, F, NREI	**0.806**	0.870	0.825
D, EL, FI, F, S	0.636	**0.961**	0.760
D, EL, F, NREI, W	0.806	0.910	**0.845**
Clues, FI, F	**0.870**	0.258	0.396
Clues, D, F, S	0.702	**0.695**	0.694
Clues, FI, F, NREI, S	0.743	0.681	**0.704**

Table 3: Precision P, recall R and f-score F of the best performing feature combinations with and without (Wilson et al., 2005)'s strong subjectivity clues.

Even though it is highly probable that some of the baseline features are either redundant, misleading, or both, no feature selection (cf. for example Weston et al. (2001)) was carried out.

3.2 Results

As we use 8 features (7 readability formula and a "feature package" consisting of 5,569 strong subjectivity clues), there are $\sum_{k=1}^{8} \binom{8}{k} = 255$ possible feature combinations. For each feature combination a SVM was trained and tested in a 10-fold cross validation setting.

In this paper we only report on the best performing feature combinations regarding precision, recall and f-score plus each single feature on its own. The results of all feature combinations will be made accessible through the author's web site[5] by the time of the publication of this work.

Results of single features are shown in Table 2, results for best performing feature combinations are shown in Table 3.

[4] http://www.cs.cornell.edu/People/pabo/movie-review-data/

[5] http://www.asv.informatik.uni-leipzig.de/staff/Robert_Remus

Formula	Minimum		Maximum		Average			Std.-Dev.	
	Subjective	***Objective***	***Subjective***	***Objective***	***Subjective***	***Objective***	***Difference***	***Subjective***	***Objective***
D	-1.86	1.06	10.81	16.43	4.62	5.06	9.58%	1.94	2.25
EL	0.0	0.00	22.0	35.0	7.12	8.15	**14.38%**	3.81	4.46
FI	3.27	3.35	7.87	10.14	4.78	5.13	7.46%	0.76	0.92
F	16.36	14.71	20.32	20.32	19.16	19.0	0.87%	0.58	0.68
NREI	-85.21	-111.79	-32.18	-33.09	-49.78	-53.82	8.18%	8.74	10.68
S	3.13	3.13	6.74	8.55	4.93	4.82	2.32%	0.68	0.79
W	7.0	7.6	75.84	114.75	31.74	36.24	14.19%	10.86	14.22

Table 4: The minima, maxima, averages and standard deviances of the presented readability formulae.

3.3 Discussion

Although the presented measures may be considered as rather crude approximations of readability, it is quite clear from the results shown in Table 2 and Table 3 that they provide a valuable source of information regarding the sentence-level subjectivity.

Whereas for "single features" the strong subjectivity clues perform best in regards to precision, every single readability formula significantly outperforms them in regards to both recall and f-score as shown in Table 2. The best performing readability measure in terms of precision is the Devereux Readability Index, the Fog Index performs best in recall and f-score. Even though these results look promising on their own, it is noteworthy that a classifier that simply always chooses the same class reaches P = 0.5, R = 1.0 and F = 0.67 on the given data set.

Combinations of different readability formulae show *considerable improvement* in precision, recall and f-score over single readability formula features as shown in Table 3. Finally, combining different readability formulae with strong subjectivity clues shows further improvement and outperforms using these clues alone in precision, recall and f-score, as also shown in Table 3.

It is remarkable that FORCAST appears in every single feature combination shown in Table 3, both with and without the subjectivity clues. Noticeably Easy Listening Formula only appears in feature combinations without the subjectivity clues. Fog Index, the best performing single readability formula, appears only in combinations including the subjectivity clues. Our own formula does not contribute a lot – it only appears once.

Comparison

(Pang and Lee, 2004) report 92% *accuracy* on sentence-level subjectivity classification using Naïve Bayes classifiers and 90% accuracy using SVMs on the same data set. (Wiebe et al., 2004) report 94% accuracy on *document-level* subjectivity classification using the *k-nearest-neighbour* algorithm. Although these results are not directly comparable to ours, our approach seems to perform not as good as theirs.

4 Conclusion & Future Work

We have shown that using readability formulae and their combinations as features in addition to already well-known subjectivity clues leads to significant quality improvements in sentence-level subjectivity classification. Therefore, one might argue in favour of our initial hypothesis and say that there is a connection between readability and subjectivity. We will carry out a detailed error analysis to shed light on their relationship.

Although our approach does not yet perform as good as current state-of-the-art, we believe that readability is a feature with *less language dependency* and a greater generalisation power than the pure presence or absence of certain word n-grams. Thus, it looks promising to further investigate in readability formulae as features for subjectivity classification. Thereby, it is possibly worthwhile to choose more complex formulae, e.g. ones that incorporate syntactic knowledge like the depth of parse trees or the number of subtrees of parse trees (cf. Schwarm and Ostendorf (2005)). Such formulae might be more reliable predictors of readability than the one used in our current work.

Questions still remaining open include: do we need readability formulae themselves or is it sufficient to just use the language characteristics captured by them? Are readability formulae independent from each other, and if so, to what degree? Are our results reproducible on other data sets, in other domains and even for languages other than English? If so, is there a plausible linguistic explanation for a correlation between subjectivity and readability? We will address these points in future work.

Acknowledgements

I'm grateful to Bo Pang and Lilian Lee, Chih-Chung Chang and Chih-Jen Lin as well as Theresa Wilson, Janyce Wiebe and Paul Hoffmann for making their data sets, their software and their lexical resources publicly available. Special thanks go to the anonymous reviewers whose useful comments and suggestions considerably improved the original paper.

References

John S. Caylor, Thomas G. Sticht, Lynn C. Fox, and J. Patrick Ford. 1973. Methodologies for Determining Reading Requirements of Military Occupational Specialties. Technical Report 73-5, HUMRO Western Division.

Chih-Chung Chang and Chih-Jen Lin, 2001. *LIBSVM: a Library for Support Vector Machines*. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector Networks. *Machine Learning*, 20(3):273–297.

Edgar Dale and Jeanne S. Chall. 1948. A Formula for Predicting Eeadability. *Educational Research Bulletin*, pages 11–28.

Irving E. Fang. 1966. The "Easy Listening Formula". *Journal of Broadcasting & Electronic Media*, 11(1):63–68.

James N. Farr, James J. Jenkins, and Donald G. Paterson. 1951. Simplification of Flesch Reading Ease Formula. *Journal of Applied Psychology*, 35(5):333–337.

Rudolf Flesch. 1944. Marks of Readable Style: A Study in Adult Education. *The Teachers College Record*, 45(6):422–423.

Robert Gunning. 1952. *The Technique of Clear Writing*. McGraw-Hill New York.

Linh Hoang, Jung-Tae Lee, Young-In Song, and Hae-Chang Rim. 2008. A Model for Evaluating the Quality of User-created Documents. In *Proceedings of the 4th Asia Information Retrieval Conference on Information Retrieval Technology*, pages 496–501.

George R. Klare. 1974. Assessing Readability. *Reading Research Quarterly*, 10(1):62–102.

Shibamouli Lahiri, Prasenjit Mitra, and Xiaofei Lu. 2011. Informality Judgment at Sentence Level and Experiments with Formality Score. pages 446–457.

Irving Lorge. 1939. Predicting Reading Difficulty of Selections for Children. *Elementary English Review*, 16(6):229–33.

Irving Lorge. 1948. The Lorge and Flesch Readability Formulae: a Correction. *School and Society*, 67:141–142.

G. Harry McLaughlin. 1969. SMOG Grading: A New Readability Formula. *Journal of Reading*, 12(8):639–646.

Hitoshi Nishikawa, Takaaki Hasegawa, Yoshihiro Matsuo, and Genichiro Kikui. 2010. Optimizing Informativeness and Readability for Sentiment Summarization. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 325–330.

Michael P. O'Mahony and Barry Smyth. 2010. The Readability of Helpful Product Reviews. In *Proceedings of the 23rd International Florida Artificial Intelligence Research Society Conference (FLAIRS)*.

Bo Pang and Lillian Lee. 2004. A Sentimental Education: Sentiment Analysis using Subjectivity Summarization Based on Minimum Cuts. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Richard D. Powers, W.A. Sumner, and Bryant E. Kearl. 1958. A Recalculation of Four Adult Readability Formulas. *Journal of Educational Psychology*, 49(2):99–105.

Ellen Riloff and Janyce Wiebe. 2003. Learning Extraction Patterns for Subjective Expressions. In *Proceedings of the 8th Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 105–112.

Ellen Riloff, Janyce Wiebe, and Theresa Wilson. 2003. Learning Subjective Nouns Using Extraction Pattern Bootstrapping. In *Proceedings of the 7th Conference on Natural Language Learning at HLT-NAACL*, pages 25–32.

Sarah E. Schwarm and Mari Ostendorf. 2005. Reading Level Assessment using Support Vector Machines and Statistical Language Models. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL)*, pages 523–530.

Edgar A. Smith. 1961. Devereux Readability Index. *The Journal of Educational Research*, 54(8):298–303.

Vladimir Vapnik. 1995. *The Nature of Statistical Learning*. Springer New York, NY.

Jason Weston, Sayan Mukherjee, Olivier Chapelle, Massimiliano Pontil, Tomaso Poggio, and Vladimir Vapnik. 2001. Feature Selection For SVMs. *Advances in Neural Information Processing Systems*, pages 668–674.

Janyce Wiebe and Ellen Riloff. 2005. Creating Subjective and Objective Sentence Classifiers from Unannotated Texts. In *Proceedings of the Sixth International Conference on Intelligent Text Processing and Computational Linguistics (CICLing)*, pages 486–497.

Janyce Wiebe, Theresa Wilson, and Matthew Bell. 2001. Identifying Collocations for Recognizing Opinions. In *Proceedings of the ACL Workshop on Collocation: Computational Extraction, Analysis, and Exploitation*, pages 24–31.

Janyce Wiebe, Eric Breck, Chris Buckley, Claire Cardie, Paul Davis, Bruce Fraser, Diane Litman, David Pierce, Ellen Riloff, Theresa Wilson, David Day, and Mark Maybury. 2003. Recognizing and Organizing Opinions Expressed in the World Press. In *Working Notes – New Directions in Question Answering (AAAI Spring Symposium Series).*

Janyce Wiebe, Theresa Wilson, Rebecca Bruce, Matthew Bell, and Melanie Martin. 2004. Learning Subjective Language. *Computational Linguistics*, 30(3):277–308.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating Expressions of Opinions and Emotions in Language. *Language Resources and Evaluation*, 1(2):165–210.

Janyce Wiebe. 2000. Learning Subjective Adjectives from Corpora. In *Proceedings of the National Conference on Artificial Intelligence*, pages 735–741.

Theresa Wilson, Janyce Wiebe, and Rebecca Hwa. 2004. Just How Mad Are You? Finding Strong and Weak Opinion Clauses. In *Proceedings of the National Conference on Artificial Intelligence*, pages 761–769.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing Contextual Polarity in Phrase-level Sentiment Analysis. In *Proceedings of the Conference on Human Language Technology (HLT) and Empirical Methods in Natural Language Processing (EMNLP)*, pages 347–354.

Hong Yu and Vasileios Hatzivassiloglou. 2003. Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences. In *Proceedings of the 8th Conference on Empirical Methods in Natural Language Processing (EMNLP)*, volume 3, pages 129–136.