Sentiment Classification of the Slovenian News Texts

Jože Bučar[1], Janez Povh[1], and Martin Žnidaršič[2]

[1] Faculty of Information Studies, Laboratory of Data technologies,
Ulica talcev 3, SI-8000 Novo mesto, Slovenia,
{joze.bucar,janez.povh}@fis.unm.si,
WWW home page: http://datalab.fis.unm.si

[2] Jožef Stefan Institute, Department of Knowledge Technologies,
Jamova cesta 39, SI-1000 Ljubljana, Slovenia,
martin.znidarsic@ijs.si,
WWW home page: http://kt.ijs.si

Abstract. This paper deals with automatic two class document-level sentiment classification. We retrieved textual documents with political, business, economic, and financial content from five Slovenian web media. By annotating a sample of 10,427 documents we obtained a labelled corpus in the Slovenian language. Five classifiers were evaluated on this corpus: Naïve Bayes Multinomial, Support Vector Machines, Random Forest, k-Nearest Neighbour and Naïve Bayes, out of which the first three were used also in the assessment of the pre-processing options. Among the selected classifiers Naïve Bayes Multinomial outperforms the Naïve Bayes, k-Nearest Neighbour, Random Forest, and Support Vector Machines classifier in terms of classification accuracy. The best selection of pre-processing options achieves more than 95% classification accuracy with Naïve Bayes Multinomial and more than 85% with Support Vector Machines and Random Forest classifier.

Keywords: sentiment analysis, document classification, machine learning, Slovenian language, corpus

1 Introduction

The unprecedented growth of the Web in the past two decades has created an entirely new way to retrieve and share information. Between 2000 and 2014, the number of web users increased more than seven times, and already surpassed three million in June 2014, which represents more than 42% of the world's population [10]. Web is the largest publicly available data source in the world and there is not enough human resources to inspect this data. People strive to detect and obtain relevant information from this chaotic cluster of data, with a hope to better understand our world and the quality of our lives.

An increasing number of blogs, newsgroups, forums, chat rooms, and social networks attracts web users to share their feelings and ideas about products, services, events, etc. The popularity of social media has escalated the interest in sentiment analysis [23]. Sentiment analysis, also known as opinion mining, is an emerging area of research for efficient analysis of informal, subjective, opinionated web content in source materials by applying natural language processing (NLP), computational linguistics and text analytics [17]. Sentiment classification is a supervised learning problem with usually three classes: negative, neutral, and positive opinion or sentiment. Its commercial use can be noticed when:

- analysing consumer habits, behaviour, trends, competitors, and market buzz,
- carrying out quality control to prevent negative viral effects,
- estimating and evaluating responses to company-related events and incidents in multiple languages,
- collecting and extracting user opinions of products and services.

When classifying documents, we deal with: document representation, feature selection, and document modelling [9]. In the first two phases we choose a feature set to represent a document. The whole content of a document is then represented with a document feature vector. In the last phase we build a document model where we use a selection of features.

In our research, we explore which of the existing and frequently used classifiers based on our data set achieves the best results in terms of time consumption and performance. Also, we study the impact of various pre-processing options on classification performance.

The rest of the paper is organized as follows: Section 2 briefly introduces the background and related work. In Section 3 we describe a proposed approach and implemented solutions. Section 4 presents the evaluation of our experimental results. Finally, we discuss the obtained results and present our prospects in section Conclusions and Future Work.

2 Background and Related Work

The first studies in sentiment document classification were made around the year 2000; even a bit earlier. They were mainly focused on financial news [5] [8], movie [18] and product reviews [21], especially on the popular Internet Movie Database (IMDb) and product reviews downloaded from Amazon [6].

McCallum and Nigam performed a comparison of event models for Naïve Bayes (NB) text classification and different vocabulary sizes on the Newsgroups data set [15]. The Naïve Bayes Multinomial (NBM) achieved 86% accuracy.

Pang et al. [18] compared the performance of various classifiers when determining the sentiment of a document, and found out that Support Vector Machines (SVM) yielded the best results in most cases. When selecting features they tried unigrams, bigrams, part-of-speech (POS) tags, and term positions, however they produced the best results using unigrams alone (82.9% accuracy).

Godbole et al. [7] worked on sources like newspapers and blog posts at the level of words, and achieved accuracy 82.7-95.7%, while others, who worked on documents, such as blog or twitter posts [19], and full web pages, have in general accuracy of around 65-85%. Despite the fact that we may be able to build comprehensive lexicons of sentiment-annotated words, there is still an issue how to detect it in a given text correctly.

Data scientists use several tools and methods to determine and classify the sentiment of digital text automatically. Sentiment classification has been studied by numerous researchers subsequently and might be the most widely studied problem in the field of sentiment analysis (see a survey in [16]). Most techniques apply supervised learning where a bag of individual words (unigrams) is the most commonly used documents representation. Large set of features have been tried by researchers such as Terms frequency (TF), Term Frequency–Inverse Document Frequency (TF-IDF) weighting schemes, POS tags, opinion words and phrases, negations, syntactic dependency [14].

In the 80's an article was published which found out that the average amount of negative news on ABC, CBS and NBC was 46.8% [20]. Since then the proportion of negative news has increased in most media. Negative news is often cheap and easy to produce, moreover, it makes profit to the media. Some media are obligated to regulate the proportion of positive and negative news.

More than 55% of all the web pages, whose content language is known, are in English. The Slovenian language is with 0.1% on the 35th place among the most common languages on the Web [22]. This is the main reason for the lack of research on sentiment classification methods suited to the Slovenian language.

3 Methodology

We retrieved 198,184 textual documents such as news articles from the digital archive of five different Slovenian web media (Žurnal24, Rtvslo, 24ur, Dnevnik, and Finance). All retrieved documents were enriched with political, business, economic and finanical content and were published between 1st September 2007 and 31st December 2013.

Initially, we removed spelling mistakes in textual content. This was followed by the annotation process where six annotators manually annotated a random sample of 10,427 documents independently (approximately 2,000 documents per web media). All annotators are native speakers and were told to specify sentiment from the perspective of an average Slovenian web user. We used a five-level Likert scale [13], in which a sentiment was given a number from 1 to 5 (1–very negative, 2–negative, 3–neutral, 4–positive, and 5–very positive). Annotation of documents was carried out on three levels independently, i.e. document-level, paragraph-level, and sentence-level. To evaluate the process of annotation Spearman correlation and Cronbach's alpha [4] were calculated. This labelled corpus was used as a training set to train, test and evaluate the classification techniques.

3.1 Sentiment Analysis Algorithm Selection

The classification of texts is one of the key tasks in text mining. The automation of procedures for the purpose of classification of texts has thus become an important activity, which has contributed to more efficient work. Data miners use a variety of tools and a wide range of learning algorithms [21] to tackle this problem. We carried out a classification and performance assessment of machine learning algorithms by applying the stratified ten-fold cross-validation method. In order to efficiently predict the category within two-class (negative and positive) document-level sentiment classification of 5,002 documents we chose to test the performance of the following classification algorithms: NB [12], NBM [15], SVM [3], k-Nearest neighbour (KNN) [1], and Random Forest (RF) [2]. The performance of selected classifiers was measured using the accuracy:

$$Accuracy = \frac{Number\ of\ Correctly\ Classified\ Examples}{Number\ of\ All\ Examples} \tag{1}$$

3.2 Pre-processing Options

Data pre-processing is an important step in sentiment analysis. We apply standard text processing techniques that include text tokenization (splitting text into individual words/terms), upper-case letters replacement with lower-case, stop word removal (removing words that do not hold relevant information), and n-gram construction (concatenating 1 to n words appearing consecutively in a document). Various pre-processing options that were used in our experiment:

- Terms with minimum frequency 2 (to eliminate terms that appear only once),
- TF or TF-IDF weighting scheme,
- Optional: replace upper-case letters with lower-case,
- Optional: stop words removal (stop words list contains nearly 1800 words),
- N-gram tokenizer (unigrams, bigrams, trigrams, and combinations).

4 Experiments and Evaluation

In the experiment six annotators labelled 10,427 documents independently, which took almost a whole year. We calculated Cronbach's alpha and Spearman coefficients (see Table 1) to describe internal consistency of annotation process. Cronbach's alpha (0.87) indicates an excellent internal consistency. Spearman coeficients also show good correlations of annotation. The maximum value of the Spearman coefficients (0.74) was found between annotator 4 and annotator 6, while the minimum was found between annotator 3 and annotator 5.

4.1 Corpora

On the basis of averaging annotations we labelled 5,425 (52%) documents as neutral, 3,337 (32%) as negative, and 1,667 (16%) documents as positive. A negative sentiment was assigned to an article if its average score was less than or equal to 2.4, neutral if its average score was greater than 2.4 and less than 3.6, and positive sentiment if its average score was greater than or equal to 3.6.

Table 1. Spearman coefficients between annotators

	Ann1	Ann2	Ann3	Ann4	Ann5	Ann6
Ann1	1	0.72	0.58	0.65	0.58	0.70
Ann2	0.72	1	0.57	0.62	0.60	0.69
Ann3	0.58	0.57	1	0.55	0.54	0.62
Ann4	0.65	0.62	0.55	1	0.61	0.74
Ann5	0.58	0.60	0.54	0.61	1	0.66
Ann6	0.70	0.69	0.62	0.74	0.66	1

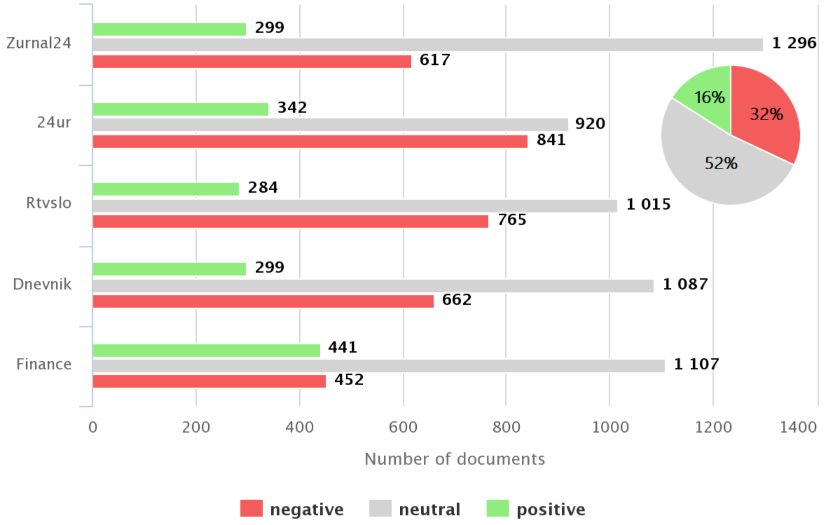

Fig. 1. Sentiment proportion per web media

We can notice that 24ur publishes the biggest proportion of negative news per medium, while Finance publishes the most positive news articles. Interestingly, the proportion of negative news is twice as large as positive in all web media, with the exception of Finance, which seems to have a more balanced content.

We generated two corpora, the first containing labelled documents as either negative or positive, while the other contains neutral documents as well. Table 2 shows statistic information about corpora.

Table 2. Corpora statistic information

Corpus	neg & pos	neg & neu & pos
# instances:	5,002	10,427
# words:	1,486,430	3,142,877
# unique words:	94,770	132,658

4.2 Document classification using various pre-processing options

In Section 3.2 we discussed several pre-processing options. In the first experiment we applied standard text preprocessing (text tokenization, stop word removal, upper-case to lower-case letters transformation, unigram, bigram, and trigram construction) on corpus with negative and positive documents (5002 instances). The resulting terms were used as features in the construction of feature vectors representing the documents, where feature vector construction was based on the TF-IDF feature weighting scheme. We also added the condition that a given term has to appear at least twice in the entire corpus. We evaluated performance using five classifiers (NB, NBM, SVM, KNN, and RF) by applying the stratified ten-fold cross-validation method. In these experiments, we achieved the accuracies of 91.84% for NBM, 86.71% for SVM, 85.19% for Random Forest with 100 trees (RF-100), 77.57% for NB, and 63.35% for k-Nearest neighbour with k=10 neighbours (KNN-10) on the test set. We also tuned two parameters:

- number of nearest neighbours at k-NN, where we applied k=1 (KNN-1), k=5 (KNN-5), and k=10 (KNN-10) nearest neighbours,
- number of trees at Random Forest classifier, where we applied 10 (RF-10), 50 (RF-50), and 100 trees (RF-100).

Table 3. Evaluation performance and time taken to train and test models for various classifiers by applying the stratified ten-fold cross-validation method

Classifier	NB	NBM	SVM	KNN-1	KNN-5	KNN-10	RF-10	RF-50	RF-100
Accuracy [%]	77.57	91.84	86.71	60.24	61.38	63.35	79.69	84.59	85.19
Time [s]	6.55	0.01	28.16	4.04	4.21	4.34	8.62	43.88	86.41

Our results show that the NBM classifier outperforms other classifiers significantly from the perspective of the classification accuracy and time taken to train and test model. SVM and RF-100 produce satisfactory results, while NB and KNN perform much worse. We also showed that tuning the number of nearest neighbours at k-Nearest Neighbour classifier and number of trees at Random Forest classifier affects the evaluation performance.

The next experiment studies which combination of the discussed pre-processing options is the best one in terms of the classification accuracy. Forty combinations of pre-processing options were tested using three classifiers (NBM, SVM, and RF-100) on corpus with negative and positive documents (5002 instances). We applied the stratified ten-fold cross-validation method. In each iteration of cross-validation, we trained classifiers on the same training partition of the data and then evaluated them on the same test partition of the data. Special care was taken not to observe the testing partition even in the pre-processing processes, thus the pre-processing is always conducted only on the learning partition of the data in each iteration of the stratified ten-fold cross-validation method. Comparisons between the applied algorithms were performed using t-paired tests with significant level set at 95%.

Table 4. Evaluation performance (accuracy and standard deviation) for NBM, SVM, and RF-100 using various pre-precessing options by applying the stratified ten-fold cross-validation method

	Pre-processing options				Avg. acc. ± std. dev. [%]				
ID	TF & TF-IDF	lower-case	stop-words	Ngrams (1,2,3)	NBM		SVM		RF-100
1	TF			1	89.08 ± 1.26	-	84.87 ± 1.07	-	82.17 ± 1.84
2	TF-IDF			1	91.40 ± 1.62	-	84.87 ± 1.07	-	81.93 ± 1.45
3	TF			2	92.06 ± 1.04	-	87.09 ± 1.50	-	80.23 ± 1.20
4	TF-IDF			2	**95.10 ± 0.92**	-	87.09 ± 1.50	-	80.05 ± 1.19
5	TF			3	89.44 ± 1.00	-	86.63 ± 1.68	-	80.51 ± 1.62
6	TF-IDF			3	93.82 ± 0.91	-	86.63 ± 1.67	-	80.33 ± 1.55
7	TF			1+2	89.18 ± 0.95	-	85.84 ± 1.72	-	82.91 ± 1.42
8	TF-IDF			1+2	91.78 ± 1.06	-	85.84 ± 1.72	-	82.39 ± 1.15
9	TF			1+2+3	89.40 ± 0.59	-	86.19 ± 1.94	-	83.19 ± 1.50
10	TF-IDF			1+2+3	91.98 ± 0.67	-	86.19 ± 1.94	-	83.17 ± 1.47
11	TF	x		1	89.34 ± 1.64	-	85.17 ± 1.69	-	82.51 ± 2.10
12	TF-IDF	x		1	91.12 ± 1.23	-	85.17 ± 1.69	-	82.45 ± 1.52
13	TF	x		2	92.08 ± 0.84	-	86.33 ± 1.14	-	80.97 ± 1.57
14	TF-IDF	x		2	94.60 ± 0.98	-	86.33 ± 1.14	-	80.57 ± 1.02
15	TF	x		3	89.34 ± 0.82	-	**87.54 ± 0.87**	-	81.03 ± 1.47
16	TF-IDF	x		3	94.10 ± 0.85	-	**87.54 ± 0.87**	-	81.07 ± 1.37
17	TF	x		1+2	89.24 ± 1.11	-	86.53 ± 1.41	-	83.49 ± 1.28
18	TF-IDF	x		1+2	91.84 ± 0.93	-	86.53 ± 1.41	-	82.87 ± 1.53
19	TF	x		1+2+3	88.96 ± 0.87	-	86.54 ± 1.65	-	83.55 ± 1.69
20	TF-IDF	x		1+2+3	91.54 ± 0.85	-	86.54 ± 1.65	-	83.59 ± 2.13
21	TF		x	1	89.06 ± 1.47	-	85.01 ± 1.52	-	84.73 ± 1.67
22	TF-IDF		x	1	91.36 ± 1.65	-	85.01 ± 1.52	-	84.99 ± 1.82
23	TF		x	2	92.06 ± 1.04	-	87.09 ± 1.50	-	80.23 ± 1.20
24	TF-IDF		x	2	**95.10 ± 0.92**	-	87.09 ± 1.50	-	80.05 ± 1.19
25	TF		x	3	89.44 ± 1.00	-	86.63 ± 1.67	-	80.51 ± 1.62
26	TF-IDF		x	3	93.82 ± 0.91	-	86.63 ± 1.67	-	80.33 ± 1.55
27	TF		x	1+2	89.78 ± 0.95	-	85.77 ± 2.11	-	84.55 ± 1.50
28	TF-IDF		x	1+2	92.22 ± 1.02	-	85.77 ± 2.11	-	84.41 ± 1.81
29	TF		x	1+2+3	89.88 ± 0.56	-	86.68 ± 1.91	-	84.59 ± 1.86
30	TF-IDF		x	1+2+3	92.26 ± 0.55	-	86.68 ± 1.91	-	84.87 ± 1.64
31	TF	x	x	1	89.22 ± 1.41	-	85.25 ± 1.47	-	84.55 ± 1.43
32	TF-IDF	x	x	1	91.50 ± 1.40	-	85.25 ± 1.48	-	84.93 ± 1.65
33	TF	x	x	2	92.08 ± 0.84	-	86.33 ± 1.14	-	80.97 ± 1.57
34	TF-IDF	x	x	2	94.60 ± 0.98	-	86.33 ± 1.14	-	80.57 ± 1.02
35	TF	x	x	3	89.34 ± 0.82	-	**87.54 ± 0.87**	-	81.03 ± 1.47
36	TF-IDF	x	x	3	94.10 ± 0.85	-	**87.54 ± 0.87**	-	81.07 ± 1.37
37	TF	x	x	1+2	90.02 ± 0.92	-	86.73 ± 1.77	-	84.73 ± 1.52
38	TF-IDF	x	x	1+2	92.04 ± 0.89	-	86.73 ± 1.77	-	84.69 ± 1.67
39	TF	x	x	1+2+3	89.50 ± 1.20	-	86.71 ± 1.79	-	84.39 ± 1.76
40	TF-IDF	x	x	1+2+3	91.84 ± 0.87	-	86.71 ± 1.79	-	**85.19 ± 1.13**

We present accuracies for all the tested pre-processing options and three classifiers (NBM, SVM, and RF-100) in Table 4. The best accuracy is obtained by NBM when using bigrams and TF-IDF scheme. Results of NBM classifier confirm that standard TF-IDF approach to feature vector construction performs better than TF weighting scheme [11]. Moreover, using bigrams and the combinations of unigrams, bigrams, and trigrams return best results. SVM performs best when applying the combinations of unigrams, bigrams, and trigrams together. It can be noticed that stop words removal does not contribute to a better performance as well as transforming upper-case letters to lower-case. RF-100 performs best when applying TF-IDF scheme, removing stop words, transforming upper-case to lower-case letters, and constructing unigrams, bigrams, and trigrams.

Relative to each pre-processing option and classifier, + (-) sign in the first column means that the accuracy of this classifier is significantly better (worse) than NBM. However, these results show, once more, that NBM is better overall.

5 Conclusions and Future Work

In this paper we introduced an automatic two class (negative and positive) document-level sentiment classification on the obtained corpus of 5,002 textual documents with political, business, economic, and financial content published between 1st September 2007 and 31st December 2013 in five Slovenian web media. This corpus is useful for both academia and industry, however it will be publicly available under CC BY license for further research. We evaluated five classifiers, especially NBM, SVM, and RF-100 which were evaluated using various pre-processing options. We also showed that the NBM classifier outperforms other classifiers significantly in terms of the classification accuracy as well as time that is needed to train and test a model.

Considering two class document-level sentiment classification we are also interested in three class (negative, neutral, and positive) document-level sentiment classification. The annotation of documents was carried out on three levels (document, paragraph, and sentence) independently, therefore we are eager to try whether fragmentation on paragraph and sentence-level can contribute to achieve a better performance. The next perspective is to study negation and how we can detect and incorporate it into the set of features. In the future, we will explore the use of feature selection, and we will apply other feature selection methods to create a sentiment lexicon for the Slovenian language.

Acknowledgements. Work supported by Creative Core FISNM-3330-13-500033 'Simulations' project funded by the European Union, The European Regional Development Fund and Young Researcher Programme by Slovenian Research Agency. The operation is carried out within the framework of the Operational Programme for Strengthening Regional Development Potentials for the period 2007-2013, Development Priority 1: Competitiveness and research excellence, Priority Guideline 1.1: Improving the competitive skills and re-search excellence.

References

1. Aha, D.W., Kibler, D., Albert, M.A.: Instance-based learning algorithms. Machine learning 6, 37–66 (1991)
2. Breiman, L.: Random forests. Machine learning 45, 5–32 (2001)
3. Cortes, C., Vapnik, V.: Support vector networks. Machine learning 20, 273–297 (1995)
4. Cronbach, L.J.: Coefficient alpha and the internal structure of tests. Psychometrika 16, 297–334 (1951)
5. Das, S.R., Chen, M.Y.: Yahoo! for amazon: Extracting market sentiment from stock message boards. In Proceedings of the Asia Pacific Finance Association Annual Conference (APFA) (2001)
6. Dave, K., Lawrence, S., Pennock, D.M.: Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In Proceedings of the World Wide Web Conference (2003)
7. Godbole, N., Srinivasaiah, M., Skiena, S.: Large-scale sentiment analysis for news and blogs. In Proceedings of the International Conference on Weblogs and Social Media 2, 1–4 (2007)
8. Hatzivassiloglou, V., McKeown, K.R.: Predicting the semantic orientation of adjectives. In Proceedings of the 8th Conference on European Chapter of the Association for Computational Linguistics, 174–181 (1997)
9. Hrala, M., and Král, P.: Evaluation of the document classification approaches. In Proceedings of the 8th International Conference on Computer Recognition Systems CORES, 877–885 (2013)
10. Internet World Stats, World Internet Users and 2014 Population Stats, 2014, `http://www.internetworldstats.com/stats.htm` 2015-03-10
11. Joachims, T.: Text categorization with support vector machines: Learning with many relevant features. In Proceedings of the European Conference on Machine Learning, 137–142 (1998)
12. Lewis, D.D.: Naïve (Bayes) at Forty: The Independent Assumption in Information Retrieval. Machine Learning: ECML-98, 4–15 (1998)
13. Likert, R.: A Technique for the Measurement of Attitudes. Archives of Psychology 22, 1–55 (1932)
14. Liu, B.: Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data. Springer-Verlag, New York, Inc., 469–492 (2011)
15. McCallum, A., Nigam, K.: A Comparison of Event Models for Naïve Bayes Text Classification. AAAI-98 workshop on learning for text categorization 752, 41–48 (1998)
16. Paliouras, G., Papatheodorou, C., Karkaletsis, V., Spyropoulos, C.: Discovering user communities on the Internet using unsupervised machine learning techniques. Interacting with Computers 14, 761–791 (2002)
17. Pang, B., Lee, L.: Opinion mining and sentiment analysis. Foundation and Trends in Information Retrieval 2, 1-135pp. 1–135 (2008)
18. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up?: Sentiment classification using machine learning techniques. In Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing, 79–86 (2002)
19. Smailović, J., Grčar, M., Lavrač, N., Žnidaršič, M: Predictive sentiment analysis of tweets: A stock market application. In Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data, 77–88 (2013)

20. Stone, G.C., Grusin, E.: Network TV as the Bad News Bearer. Journalism Quarterly 61, 517 (1984)
21. Turney, P.D.: Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. Proceedings of the 40th annual meeting on association for computational linguistics, Association for Computational Linguistics, 417–424 (2002)
22. Web Technology Surveys Usage of content languages for websites, 2011, http://w3techs.com/technologies/overview/content_language/all 2015-03-08
23. Wright, A.: Mining the Web for Feelings, Not Facts. New York Times 24 (2009)