2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology

Opinion Mining on Newspaper Quotations

Alexandra Balahur, Ralf Steinberger, Erik van der Goot, Bruno Pouliquen, Mijail Kabadjov
European Commission - Joint Research Centre
{Alexandra.Balahur-Dobrescu, Ralf.Steinberger, Erik.van-der-Goot, Bruno.Pouliquen, Mijail.Kabadjov}@jrc.ec.europa.eu

Abstract— **Opinion mining is the task of extracting from a set of documents opinions expressed by a source on a specified target. This article presents a comparative study on the methods and resources that can be employed for mining opinions from quotations (reported speech) in newspaper articles. We show the difficulty of this task, motivated by the presence of different possible targets and the large variety of affect phenomena that quotes contain. We evaluate our approaches using annotated quotations extracted from news provided by the EMM news gathering engine. We conclude that a generic opinion mining system requires both the use of large lexicons, as well as specialised training and testing data.**

Keywords- opinion mining; sentiment analysis; media monitoring.

I. INTRODUCTION

Many news monitoring systems are available on the internet, including NewsVine[1], NewsTin[2] and Europe Media Monitor[3]. Their functionality may include breaking news alerting, categorisation of news according to pre-defined categories or user-defined search words, linking of related news over time and across languages, extraction and display of meta-information such as references to locations, persons and organisations, or quotations. Another important aspect media analysts are interested in is sentiment, or opinion. Opinion mining applications try to detect news bias across sources, subjectivity of reporting; they try to measure the popularity of persons, organisations, products, programs, and more.

In this paper, we present ongoing work on adding opinion mining to the freely accessible Europe Media Monitor (EMM) family of media analysis applications. We are specifically working on detecting sentiment in quotations (direct reported speech). The reason for this is that the text in quotes is usually more subjective than the other parts of news articles, where sentiment is either expressed less, or it is expressed less explicitly. We also know for quotes who the person is that made the statement (referred to as the source of the opinion statement) and – if the speaker makes reference to another entity within the quotation – we have a clue about the possible target (or object) of the sentiment statement. (e.g. *Steinmeier said: "I think we can conclude that there is a fresh wind in NATO, and also, hopefully, a new atmosphere of cooperation,"*)

The fact that source and target are frequently known in quotations makes them also suitable for the detection of social networks from free text. [1] used automatically learned patterns to detect bi-valent relations such as support and criticism between persons and used these relations to produce a social network of persons mentioned in the news. In further work, the sentiment detected in quotations was used to expand this original social network with more positive and negative relations [2].

The objective of the work presented here is to make a preliminary assessment of methods to refine the detection and classification of sentiment or opinion in reported speech. Results will eventually be fed into EMM-NewsBrief and EMM-NewsExplorer, which display information automatically extracted from the news in 19 or more languages, including quotations by and about named entities.

II. BACKGROUND AND RELATED WORK

Opinion mining (sentiment analysis) is the task of extracting, given a set of documents, the opinion expressed in them by a *source*, on a certain *target*. While the field is relatively recent in Natural Language Processing, extensive research has already been conducted within its framework, motivated by the Web 2.0 phenomena. The main reason is that new textual genres (emerging text types) in the "Social Web", such as blogs, e-forums or e-commerce reviews contain a "snapshot" of the current opinion of people from all over the world on a high diversity of topics, from economy, politics and environment to electronics. This large volume of subjective data can be exploited for comparing opinions of other people on products (swooty.com), market feedback, monitoring of image and reputation, monitoring and analysing social media, detecting general mood (wefeelfine.org – sentiment analysis over the blogosphere) and detecting "hot" news (twends.net – sentiment analysis over the Twitter social network).

However, initial research concentrated on news texts. [3] defines subjectivity based on Quirk's idea of "private states" (states that are not open to verification) and distinguishes between objectivity and subjectivity on this criteria. Consequently, based on this definition, the Multi-Perspective Question Answering (MPQA) annotation

[1] http://www.newsvine.com
[2] http://www.newstin.com
[3] http://press.jrc.it/overview.html

978-0-7695-3801-3/09 $26.00 © 2009 IEEE
DOI 10.1109/WI-IAT.2009.340

schema and corpus were created over news texts, distinguishing between subjective/objective speech, as well as the polarity of text spans [4]. Subsequently, different authors show that this initial discrimination is crucial for the sentiment task, improving results obtained when using only polarity classification for sentence-level opinion mining [5], as part of Opinion Information Retrieval (last three editions of the TREC Blog tracks, the TAC 2008 competition), Information Extraction [6] and Question Answering (QA) [7] systems. Once this discrimination is done, or in the case of texts containing only or mostly subjective language (such as e-reviews), opinion mining becomes a polarity classification task.

Research conducted in analysing sentiment at a sentence level used bootstrapping techniques [8], considered gradable adjectives [9], semi supervised learning with the initial training set identified by some strong patterns and then applying Naïve Bayes or self-training [10], focused on finding strength of opinions [11], summing up orientations of opinion words in a sentence after filtering by subjectivity, using the Web as corpora [12, 13], determining the semantic orientation of words and phrases [14], identifying opinion holders [8], comparing sentence and relation extraction and feature-based opinion mining and summarization [15].

III. MOTIVATION AND CONTRIBUTION

As mentioned in the previous section, extensive work has already been conducted on opinion mining, at different levels of text and on different polarity scales. Applications include a variety of areas, depending on the source and final user of the extracted data – from monitoring the image of public figures to company reputation or trust, monitoring and analysing social media to detect potentially dangerous situations and what is done about them, or tracking opinion across time for market and financial studies.

Relevant information, however, must be obtained from reliable sources. The news data used in our experiments is that provided by the EMM news gathering engine. EMM gathers an average of 80-100,000 news articles per day in about 50 languages, from about 2,200 mostly internet-based media sources [16]. From these, an average of 3165 quotations is extracted every day [17] for all languages, from which 1630 are for English. Only those quotations are retained where the speaker (the source) could be identified unambiguously, which are 319 in all languages and 174 for English. EMM applications are visited by some 30,000 and 50,000 users every day and they get between 1 and 2 million hits per day. While most EMM applications are highly multilingual, we are currently only experimenting with English quotations. As the objective is to eventually apply the same methods to all EMM languages, we made sure to only use simple methods for which no or very few linguistic analysis tools and resources will be needed.

News data is very different from product reviews in that sentiment is usually expressed much less explicitly. Bias or sentiment can be expressed by mentioning some facts while omitting others, or it can be presented through subtle methods like sarcasm (e.g. *"Google is good for Google, but terrible for content providers"*). We do not claim that we will detect such instances. Instead, we focus on detecting those (relatively) explicit opinion statements found in the news, and especially in quotations. Another major difference between the news and product reviews is that the target of the sentiment is much less concrete. A camera has features like weight, flash light, battery life, etc., but what are the features of a named entity such as a specific person or of an organisation like the European Commission (EC)? And how to identify these features, which may include an infinite array of things, including efficiency of the administration, various policy areas, impact on the development of poorer regions, or on consumer protection and environment issues? It is also rather tricky to detect whether any negative sentiment detected refers to the organisation EC, or to the main news content in which the EC was mentioned (e.g. a natural disaster, to which the EC may be reacting with aid and support). Instead of trying to tackle all of these complex issues, our current aim is to categorise quotations for subjectivity (neutral vs. subjective) and to determine the polarity of the subjective quotations. Unlike full articles, quotations are relatively short and often are about one subject. However, they contain a variety of interesting phenomena, such as the combination of a short, factual summary of the event or what the "target" did or a general view on the problem, as well as the opinion or position of the "source" on this fact description (e.g. *"It is a tough battle and those who perceive us as competitors are not going to roll over and play dead. But again, both Branson and Fernandes are battle scarred and, with a song and a prayer, and lots of hard work, I believe we shall prevail"*). Another interesting aspect concerns the presence of various possible "targets" in the quote, on which antonymic opinions are expressed (e.g. *"How can they have a case against Fred, when he didn't sign anything?"*). Moreover, they contain a larger scale of affective phenomena, which are not easily classifiable when using only the categories of positive and negative: warning (e.g. *"Delivering high quality education cannot be left to chance!"*), doubt (e.g. *"We don't know what we should do at this point")*, concern, confidence, justice etc. (where doubt is generally perceived as a negative sentiment and confidence as a positive one). Last, but not least, a note should be made on the interpretation of the "opinion mining" task. As noticed in [18], there are two understandings of the notion, one regarding the bad versus good *news* classification and the other regarding the positive versus negative *attitude* classification. It is important to stress that in our approach, the aim is to determine the attitude polarity (tonality of the speech), independent of the type of news, interpreting only the content of the text and not the effect it has on the reader.

IV. EXPERIMENTS AND EVALUATION

A. Data

For our experiments, we chose a set of 99 quotes, on which agreement between a minimum of two annotators could be reached regarding their classification in the positive and negative categories, as well as their being neutral/controversial or improperly extracted. The result of the grouping was a total of 35 positive, 33 negative, 27 neutral/controversial and 4 improperly extracted quotes. We used this dataset to comparatively analyse the different possible methods and resources for opinion mining and we explored the possibility to combine them in order to increase the accuracy of the classification. The first approach is based on a "bag of words" – the use of different lexicons containing positive and negative words. The second approach contemplates measuring similarity to existing corpora and machine learning.

B. Bag-of-words approach

At the present moment, there are different lexicons for affect detection and opinion mining. The aim in the following evaluation is to test the different resources in the quote classification scenario and assess the quality and consistency of these lexicons. Each of the employed resources were mapped to four categories, which were given different scores – positive (1), negative (-1), high positive (4) and high negative (-4). The assignment of these values was based on the intuition that certain words carried a higher affective charge and their presence should be scored accordingly. Our intuition was supported by experiments in which we used just the positive and negative categories and that scored lower. The polarity value of each of the quotes was computed as sum of the values of the words identified; a positive score leads to the classification of the quote as positive, whereas a final negative score leads to the system classifying the quote as negative. The resources used were: the JRC lists of opinion words, WordNet Affect [19], SentiWordNet [20], MicroWNOp [21]. WordNet Affect categories of anger and disgust were grouped under high negative, fear and sadness were considered negative, joy was taken as containing positive words and surprise as highly positive; SentiWordNet and MicroWNOp contained positive and negative scores between 0 and 1 and in their case, we mapped the positive scores lower than 0.5 to the positive category, the scores higher than 0.5 to the high positive set, the negative scores lower than 0.5 to the negative category and the ones higher than 0.5 to the high negative set. As a filtering step, we first classified the quotes based on the presence of "subjectivity indicators" [22]. The subjective versus objective filtering had an accuracy of 0.89, as 2 of the positive and 5 of the negative quotes were classified as neutral. We evaluated the approaches both on the whole set of positive and negative quotes, as well as only the quotes that were classified as "subjective" by the subjectivity indicators. Subsequently, we grouped together resources that tended to over-classify quotes as positive or negative, in an attempt to balance among their classification. Finally, we grouped together all the words pertaining to the different classes of positive, negative, high positive and high negative words belonging to all the evaluated resources. The results are shown in Table 1 (-S/O and +S/O indicate absence and presence, respectively, of the subjectivity filtering):

TABLE I. RESULTS OF THE CLASSIFICATION USING THE DIFFERENT OPINION AND AFFECT LEXICONS

Resource	-S/O	+S/O	P_{pos}	P_{neg}	R_{pos}	R_{neg}
JRCLists	X		0.77	0.3	0.54	0.55
		X	0.81	0.35	0.6	0.625
SentiWN	X		1	0	0.51	0
		X	1	0	0.54	0
WNAffect	X		0	1	0	0.51
		X	0	1	0	0.54
MicroWN	X		0.62	0.36	0.52	0.48
		X	0.73	0.35	0.57	0.53
SentiWN + WNAffect	X		0.22	0.66	0.42	0.45
		X	0.24	0.67	0.47	0.41
All	X		0.68	0.64	0.7	0.62
		X	0.73	0.71	0.75	0.69

C. Similarity approach

In this approach we used two existing resources – the ISEAR corpus [23] - consisting of phrases where people describe a situation when they felt a certain emotion and EmotiBlog [24], a corpus of blog posts annotated at different levels of granularity (words, phrases, sentences etc.) according to their polarity and emotion.

In the first approach, we computed the individual quotes' similarity with the sentences belonging to each of the emotions in the ISEAR corpus, using Pedersen's Similarity Package[4], based on the Lesk similarity[5]. Subsequently, we classified each of the quotes based on the highest-scoring category of emotion. Table 2 presents the results:

TABLE II. RESULTS OF THE CLASSIFICATION USING THE SIMILARITY SCORES WITH THE ISEAR CORPUS

Class	Joy	Fear	Anger	Shame	Dis-gust	Guilt	Sad-ness
Positive	8	7	1	3	3	5	8
Negative	6	7	1	5	8	2	4

We consider as positive the examples which fell into the "joy" category and classify as negative the quotes which were labelled otherwise. The results are presented in Table 3:

TABLE III. RESULTS OF THE POSITIVE VERSUS NEGATIVE CLASSIFICATION USING THE SIMILARITY SCORE WITH THE ISEAR CORPUS

Ppos	Pneg	Rpos	Rneg	Accuracy
0.22	0.82	0.58	0.5	0.514

[4] http://www.d.umn.edu/~tpederse/text-similarity.html

[5] http://kobesearch.cpan.org/htdocs/WordNet-similarity/WordNet/Similarity/lesk.htm

EmotiBlog represents an annotation schema for opinion in blogs and the annotated corpus of blog posts that resulted when applying the schema. The results of the labelling were used to create a training model for an SVM classifier that will subsequently be used for the classification of opinion sentences. The features considered are the number of n-grams (n ranging from 1 to 4) and similarity scores with positive and negative annotated phrases, computed with Pedersen's Similarity Package. The approach is described in [25]. The evaluation results are presented in Table 4:

TABLE IV. RESULTS OF THE CLASSIFICATION USING SVM ON THE EMOTIBLOG CORPUS MODEL

Class	Precision	Recall	F-measure
Positive	0.667	0.219	0.33
Negative	0.533	0.89	0.667

V. DISCUSSION, CONCLUSIONS AND FUTURE WORK

From Table 1 we can infer that the use of some of the resources leads to better performance when classifying positive or negative quotes (SentiWN versus WordNet Affect), and that the combined resources produce the best results when a vocabulary-based approach is used. Another conclusion is that previous subjectivity filtering indeed improves the results. Regarding the emotion categories, Table 3 shows that precision is higher for negative emotions, presumably because there is only one positive emotion class (joy). It would thus be useful to add more positive classes. Finally, from the results in Table 4, we can conclude that annotations about a specific topic cannot be applied to generic opinion mining on news. This confirms that open-domain opinion analysis is a more difficult problem than topic-specific sentiment classification and other sub-tasks defined in opinion mining. Experiments showed that simple bag-of-words approaches cannot reach a satisfactory level, even when large sets of words are employed. Future work includes exploiting bootstrapping methods to produce sentiment vocabulary for other languages in which EMM articles are gathered, learning patterns for opinion expression, employing methods for target/topic identification and, in the case of larger topics, considering context for opinion classification.

REFERENCES

[1] Tanev, H. "Unsupervised Learning of Social Networks from a Multiple-Source News Corpus". Proceedings of the Workshop Multi-source Multilingual Information Extraction and Summarization (MMIES'2007) held at RANLP 2007.

[2] Tanev, H., Pouliquen, B., Zavarella, B. and Steinberger, R. "Automatic expansion of a social network using sentiment analysis". In Annals of Information Systems, Special Issue on Data Mining for Social Network Data (unpublished).

[3] Wiebe, J. "Tracking point of view in narrative". Computational Linguistics 20 (2): 233-287, 1994

[4] Wiebe, J., Wilson, T. and Cardie, C. "Annotating expressions of opinions and emotions in language". Language Resources and Evaluation, volume 39, issue 2-3, pp. 165-210, 2005.

[5] Pang, B. and Lee, L. "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts". In Proceedings of the ACL 2004.

[6] Riloff, E., Wiebe, J., and Phillips, W. "Exploiting Subjectivity Classification to Improve Information Extraction". Proceedings of the 20th National Conference on Artificial Intelligence, AAAI-05.

[7] Stoyanov, V., Cardie, C. "Toward Opinion Summarization: Linking the Sources". In: COLING-ACL 2006 Workshop on Sentiment and Subjectivity in Text.

[8] Riloff, E., Wiebe, J. "Learning Extraction Patterns for Subjective Expressions". In Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing, EMNLP 2003.

[9] Hatzivassiloglou, V., Wiebe, J. "Effects of adjective orientation and gradability on sentence subjectivity." In Proceedings of COLING 2000.

[10] Wiebe, J., Riloff, E. "Creating Subjective and Objective Sentence Classifiers from Unannotated Texts". In Proceedings of the 6th International Conference on Computational Linguistics and Intelligent Text Processing , CICLing 2005.

[11] Wilson, T., Wiebe, J., Hwa, R. "Just how mad are you? Finding strong and weak opinion clauses". In: Proceedings of AAAI 2004.

[12] Kim, S.M., Hovy, E. "Determining the Sentiment of Opinions". In Proceedings of COLING 2004.

[13] Lin, W.H., Wilson, T., Wiebe, J., Hauptman, A. "Which Side are You On? Identifying Perspectives at the Document and Sentence Levels". In Proceedings of the Tenth Conference on Natural Language Learning CoNLL 2006.

[14] Turney, P., Littman, M. "Measuring praise and criticism: Inference of semantic orientation from association". ACM Transactions on Information Systems 21, 2003

[15] Turney, P. "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of review"s. In Proceedings of the 40th Annual Meeting of the ACL 2002.

[16] Atkinson, M. and Van der Goot, E. "Near Real Time Information Mining in Multilingual News." In: 18th International World Wide Web Conference, WWW 2009.

[17] Pouliquen, B., Steinberger, R., Best, C. "Automatic Detection of Quotations in Multilingual News". In Proceedings of the International Conference Recent Advances in Natural Language Processing , RANLP 2007.

[18] Pang, B. and Lee, L. "Opinion mining and sentiment analysis". In Foundations and Trends in Information Retrieval, Vol. 2, Nos. 1-2, pp. 1-135, 2008.

[19] Strapparava, C. Valitutti, A. "WordNet-Affect: an affective extension of WordNet". In Proceedings of the 4th International Conference on Language Resources and Evaluation , LREC 2004.

[20] Esuli, A., Sebastiani, F. "SentiWordNet: A Publicly Available Resource for Opinion Mining". In Proceedings of LREC 2006.

[21] Cerini, S., Compagnoni, V., Demontis, A., Formentelli, M., and Gandini, G. Language resources and linguistic theory: Typology, second language acquisition, English linguistics, chapter Micro-WNOp: A gold standard for the evaluation of automatically compiled lexical resources for opinion mining. Franco Angeli Editore, Italy.

[22] Wilson, T., Wiebe, J., Hoffman, P. "Recognizing contextual polarity in phrase-level sentiment analysis". In Proceedings of HLT-EMNLP 2005.

[23] Scherer, K. and Wallbott, H.G. |The ISEAR Questionnaire and Codebook", 1997.

[24] Balahur, A., Boldrini, E., Montoyo, A. and Martínez- Barco, P. "Fact Versus Opinion Questions Classification and Answering: Challenges and Keys". In Proceedings of the International Conference on Artificial Intelligence , ICAI 2009.

[25] Balahur, A., Boldrini, E., Montoyo, A. and Martínez- Barco, P. "Cross-topic Opinion Mining for Real-time Human-Computer Interaction". In Proceedings of ICEIS 2009.