”Yes we can?”: Subjectivity Annotation and Tagging for the Health Domain

Muhammad Abdul-Mageed
Department of Linguistics and
School of Library & Info. Science,
Indiana University,
Bloomington
mabdulma@indiana.edu

Mohammed Korayem
School of Informatics
& Computing
Indiana University,
Bloomington
mkorayem@indiana.edu

Ahmed YoussefAgha
Applied Health Science Department
Indiana University,
Bloomington
ahmyouss@indiana.edu

Abstract

The area of *Subjectivity and sentiment analysis (SSA)* has been witnessing a flurry of novel research. However, only few attempts have been made to build SSA systems for the health domain. In the current study, we report efforts to partially bridge this gap. We present a new labeled corpus of professional articles collected from major Websites focused on the Obama health reform plan (OHRP). We introduce a new annotation scheme that incorporates subjectivity as well as topics directly related to the OHRP and describe a highly-successful SSA system that exploits the annotation. In the process, we introduce a number of novel features and a wide-coverage polarity lexicon for the health domain.

1 Introduction

In recent years, searches and processing of data beyond the limiting level of surface words are becoming more important than it used to be (Diab et al., 2009). One of the areas that has been witnessing a swelling interest is that of *Subjectivity and sentiment analysis (SSA)*. *Subjectivity* in natural language refers to aspects of language used to express opinions, feelings, evaluations, and speculations (Banfield, 1982; Wiebe, 1994) and it, thus, incorporates *sentiment*. *Subjectivity classification* refers to the task of classifying texts into either *Objective* (e.g., *The Obama Health Committe submitted a report last week.*) or *Subjective*. Subjective text is further classified with *sentiment* or *polarity*. For sentiment classification, the task refers to identifying whether a subjective text is *positive* (e.g., *Obama's reform plan will solve all our health problems!*), *negative* (e.g., *The proposed ideas will lead to definite failure!*), *neutral* (e.g., *The president may make changes to some of the ideas proposed.*), and, sometimes, *mixed* (e.g., *The plan is bad, but I like Obama.*).

In spite of the great interest in SSA, only few studies have been conducted on the health domain. The quick dissemination of information characteristic of our world today, makes opinions expressed in these media more important than they traditionally used to be, and hence building SSA systems on top of these media is a valuable endeavor. In the current paper, we present a paragraph-level novel annotation scheme for professional articles from the health domain that incorporates customized topic annotation. More specifically, we focus on articles treating the Obama Healthcare Reform Plan (OHRP).

The rest of the paper is organized as follows: In Section 2, we motivate work on the news genre. In Section 3, we introduce data set, summarize subjectivity and topic annotations, and provide examples of categories in our data. In Section 4 we describe our approach. In Section 5, we describe our system. In Section 6 we provide the results and evaluation. Section 7 is the about related work, and Section 8 is the conclusion.

2 Professional Articles

Most SSA work has focused on highly subjective, user-generated genres such as blogs and product or

Proceedings of Recent Advances in Natural Language Processing, pages 666–671,
Hissar, Bulgaria, 12-14 September 2011.

movie reviews where authors express their opinions quite freely (Balahur and Steinberger, 2009). Professional articles (i.e., position articles written by experts) published by major news organizations is a genre that has almost been disregarded in SSA. These articles tend to differ from regular news stories reporting events in that their authors are highly specialized. Although the sentiment expressed in regular news articles is usually subtle professional articles observably have more explicit sentiment that usually differs depending on the specific dimension of the topic under discussion. In this way, the sentiment can easily shift from a paragraph to another. For this specific reason, our annotation is fine-grained (i.e., conducted at the paragraph level).

3 Data set and Annotation

3.1 Corpus

The corpus is collection of news articles crawled from 105 popular online news sites (e.g., ABC News, The Associated Press, Belfast Telegraph)Articles were selected by searching the websites using all possible combinations of the queries ”Obama healthcare,” ”Obama health reform,” and ”health care reform”. Only articles written by professionals treating the specific subject of OHRP that were published between October 2008 and September 2010 were included. Since our unit of analysis is the paragraph, articles were divided into their component paragraphs (making up 1850 paragraphs).

3.2 Subjectivity and Sentiment Annotation

We prepared guidlines for the task of subjectivity and sentiment annotation. In the current paper we summarize some of these guidelines, and cite some of the related literature.

Subjectivity and Sentiment Categories: For each paragraph, each annotator assigned one of 5 possible labels: (1) OBJECTIVE (OBJ), (2) SUBJECTIVE-POSITIVE (S-POS), (3) SUBJECTIVE-NEGATIVE (S-NEG), (4) SUBJECTIVE-NEUTRAL (S-NEUT), and (5) SUBJECTIVE-MIXED (S-MIXED). We followed (Wiebe et al., 1999) in operationalizing the SUBJ vs. OBJ categories. In other words, if the primary goal of a paragraph is perceived to be the objective reporting of information, it was labeled OBJ. Otherwise, the paragraph would be a candidate for one of the four SUBJ classes. Two college-educated native speakers of English annotated the 1850 paragraphs for both subjectivity, with inter-rater agreement at 84%. Our data has 1571 SUBJ and 279 OBJ cases. The SUBJ cases are broken into 237 S-POS, 301 S-NEG, 707 S-NEUT, and 326 S-MIXED cases

3.3 Topic Annotation

The same two colledge-educated native speakers of English who coded the data for SSA also manually assigned each paragraph a domain/topic label. The topic labels are inspired by the Obama administration’s focus on three main topics for popularizing the OHRP: (1) stability & security, (2) quality & affordability, and (3) funding. [1]. The set of topic labels is thus as follows: {*STABILITY & SECURITY (297 cases), (2) QUALITY & AFFORDABILITY (380 cases), (3) FUNDING (328 cases), OTHER (845 cases)*}. We did not make further attempts to identify other topics outside the scope of the administration’s focus. Topic annotation turned out to be an easier task than subjectivity annotation, which is indicated by inter-annotator agreement for topic label assignment being at 94%. Explanations of each category in our data set are provided in Section 3.4, with some illustrating examples.

3.4 SSA and Topic Examples

Stability & Security: Descriptions of the *Stability & Security* topic/dimension included that the plan (1) ends discrimination against people with pre-existing conditions, (2) prevents insurance companies from dropping coverage when people are sick and need it most, etc.Below, we provide one example labeled with this topic from the OBJ class:[2]

- ”I was denied coverage as spinal fractures were misdiagnosed (by the insurer’s doctor, who avoided the cost of a CT scan) concluding my 25% spinal misalignment was pre-existing.” **(OBJ)**

Quality & Affordability: Descriptions of the *Quality & Affordability* included that the plan (1) creates

[1] www.whitehouse.gov/assets/documents/obama_plan_card.PDF

[2] For limitations of space, we are not able to provide examples belonging to all our SSA categories.

a new insurance marketplace the Exchange that allows people without insurance and small businesses to compare plans and buy insurance at competitive prices, (2) provides new tax credits to help people buy insurance and to help small businesses cover their employees, etc. The following is an example:

- ”Massachusetts became the only state to mandate health insurance in 2006. It has passed legal muster and led to 97 percent of residents having some form of coverage, said Alan Sager, director of the Health Reform Program at Boston University’s School of Public Health.” **(OBJ)**

Funding: Descriptions of the *Funding* dimension included that the plan (1) will not add a dime to the deficit and is paid for upfront, (2) creates an independent commission of doctors and medical experts to identify waste, fraud and abuse in the health care system, etc. Below is an example:

- ”The House plan is projected to guarantee coverage for 96 percent of Americans at a cost of more than $1 trillion over the next 10 years, according to the nonpartisan Congressional Budget Office.” **(OBJ)**

4 Approach

4.1 Features

The following are the set of features we apply:

TOPIC: We apply a feature indicating the *topic/dimension* of the each paragraph.

UNIQUE: Following Wiebe et al. (2004), to account for the frequency of words’ effect, we include a *unique* feature. Namely words that occur in our corpus with a frequency $\leq$3 are replaced with the token ”UNIQUE”.

N-GRAM: We run experiments with *N*-grams $\leq$3 and all possible combinations of them. Thus, we employ *N*-gram combinations, as follows:(1) 1g, (2) 2g, (3) 3g, (4) 1g+2g, (5) 1g+3g, (6) 2g+3g, (7) 1g+2g+3g.

POLARITY_LEX: We apply a binary *has_polar* feature indicating whether or not any of the polarized entries in a polarity lexicon. We compare the performance of a number of polarity lexicons, including a manually labeled lexicon we manually developed i.e., the YouTube Lexicon (YT_LEX). We describe YT_LEX as well as the other lexicons we use below:

- **YT_LEX**: We used Google’s YouTube Data API to crawl all comments on 1000 YouTube videos using the query ”obama health care”. This corpus, which we refer to as *YouTube Health Corpus [YuHC]* is harvested as part of another project we are working on and totals 229,177 comments. After reducing all repeated letters of frequency ¿ 2 to only 2 (e.g., the word *cooool* is reduced to *cool*), we extracted the top 29.991 words[3] and manually labeled them with semantic orientation tags. Each word was given a label of the set {*positive, negative, neutral*}. We refer to this lexicon as the *YT_LEX*.

- **HW_LEX**: This is a list of adjectives comprising all gradable and dynamic adjectives, both manually prepared and automatically extracted, by (Hatzivassiloglou and Wiebe, 2000)[4].

- **SentiWN_LEX**: This lexicon is composed of all positive and negative entries with a score $>$ 0.25 [5] from SentiWordnet 3.0 (Baccianella et al., 2010).

- **SentiWN_Strong_LEX**: This lexicon is composed of all positive and negative entries with a score $>$ 0.50 [6] from Sentiwordnet 3.0.

SOURCE: We apply a ”SOURCE” feature to each paragraph vector. This feature indicated the news source (i.e., the news site/organization [e.g., SOURCE_CNN, SOURCE_CNBC]) from which the paragraph’s document was collected. This feature is intended to capture any bias with or against the OHRP, or one or more aspect of it, on the part of the news site/organization.

AUTHOR: We apply an ”AUTHOR” feature to each paragraph vector. This feature indicated the author of each the document to which the paragraph belongs. Again, this feature is intended to capture

[3] Extracted words were of frequency of 3 or more.

[4] The list is made available by (Hatzivassiloglou and Wiebe, 2000) here: http://www.cs.pitt.edu/ wiebe/pubs/coling00

[5] We averaged the score for repeated entries (i.e., those with more than one sense).

[6] We also averaged the score for entries with more than one sense.

any bias with or against the OHRP, or one or more aspect of it, on the part of the author.

Both the SOURCE and AUTHOR features can be viewed as meta-data features. These two features are novel ones that we introduce to the task of paragraph-level subjectivity analysis. One advantage of these two features is that they are easy to incorporate as a document is pre-processed, and hence do not need any manual tagging.

5 Automatic tagging of Subjectivity

5.1 Method

In this study, we only report experiments for subjectivity classification where attempts are made to tease apart the SUBJ from OBJ cases in our dataset. Since our data set is very biased toward the SUBJ class, we equalize the two classes by making use of all the 279 OBJ cases and randomly sampling 279 SUBJ cases from the corpus. All experiments reported below are hence run on this equalized data sample, with a baseline of 50%.

We use an Support Vector Machine classifier SVM$^{\text{light}}$ package (Joachims, 2008). We experiment with various kernels and parameter settings and find that linear kernels yield the best performance for our specific problem. We run experiments with *presence* vectors, i.e. for each sentence vector, the value of each dimension is binary either a 1 (regardless of how many times a feature occurs) or 0.

Experimental Conditions: We run three sets of experiments. We first run experiments using each of the three features *TOPIC (T), SOURCE (S), AUTHOR (A)* separately and then combined across the various *N-GRAM* and *N-GRAM* combinations described earlier. We call this first set of experiments TSA_EXP. Second, we run the UNIQUE_EXP experiments where we apply the "UNIQUE" feature explained earlier with the best-yielding *N-GRAM* or *N-GRAM* combination from TSA_EXP. Third, we run the POLAR_EXP experiments using each of the polarity lexicons separately with the following configurations: (1) the best yielding *N-GRAM* or *N-GRAM* combination from TSA_EXP, (2) the best-yielding feature (i.e., TOPIC, SOURCE, or AUTHOR) or feature combination (TOPIC+SOURCE+AUTHOR) from TSA_EXP, (3) the best yielding setting from UNIQUE_EXP, and (4) the combination of 3 and 4 configurations (i.e., the best-yileding feature or feature combination from TSA_EXP and the best-yielding setting from UNIQUE_EXP).

6 Results and Evaluation

We report results in 10-fold cross validation where we train on 9 folds and test on the 10th and average the results. Results are reported in accuracy A and F-measure (F).

TSA_EXP: As table 1 shows, each of the three features TOPIC, SOURCE, and AUTHOR improves the classification when applied. For the TOPIC feature, whereas the best A is 72.21% and is acquired using unigrams (i.e., 1g), the best F is 73.76% and is achieved with the unigram+bigram (i.e., 1g+2g) combination. Although these results are slightly higher than the results acquired using only the bag-of-words, they are $>$ 20.00% better than the 50.00% majority class baseline. Using the SOURCE feature resuls in 88.51% A and 87.93% F with bigrams, and hence an improvement of 24.24% A and 18.30% F over the results acquired with the bag-of-words with bigrams. Better results are, however, acquired when the AUTHOR feature is applied, with A reaching 95.50% and F reaching 95.51%. Applying the three features TOPIC, SOURCE, and AUTHOR combined results 95.15% A and 94.97% F. In this way, applying the AUTHOR feature alone achieves the best permofrmance.

UNIQUE_EXP: Since the best performance (in both A and F) from TSA_ EXP was with trigrams, we apply the UNIQUE feature with the trigram configuration. As table 2 below shows, we apply the UNIQUE feature with the number of words replaced by the "UNIQUE" token $\leq$ 5 absolute frequency. We acquire the best results when we replace tokens with frequency =3, with 60.43% A and 68.91% F. This is an improvement of 10.43% A and 18.90% F over the baseline.

POLAR_EXP: As stated earlier, POLAR_EXP experiments were run with four different configuration. The four configurations are (1) BASE TRIGRAMS (i.e., only trigrams), (2) BASE TRIGRAMS+UNIQUE3 (i.e., the UNIQUE feature with frequency =3), (3) BASE TRIGRAMS+AUTHOR, and (4) BASE TRI-

N-gram	Bag-of-Words		Topic		Source		Author		Topic+Source+Author	
	A	F	A	F	A	F	A	F	A	F
1g	**70.42**	72.11	**72.21**	73.50	85.82	85.79	86.54	86.80	93.35	93.19
2g	64.27	69.63	68.96	72.74	**88.51**	**87.93**	93.17	93.36	**95.15**	**94.97**
3g	54.66	67.93	63.51	67.99	86.88	86.30	**95.50**	**95.51**	95.15	94.92
1g+2g	70.06	72.70	71.48	73.76	82.60	82.97	79.37	80.69	89.59	89.71
1g+3g	68.81	71.22	69.51	72.10	82.95	83.28	79.73	81.21	90.12	90.10
2g+3g	60.86	69.07	64.83	70.55	87.44	87.16	89.40	90.33	93.36	93.11
1g+2g+3g	68.44	71.55	70.41	73.59	79.37	80.47	76.33	78.52	86.36	86.77
Baseline	50%		50%		50%		50%		50%	

Table 1: TSA_EXP Results

N-gram	Bag-of-Words		unique1		unique2		unique3		unique4		unique5	
	A	F	A	F	A	F	A	F	A	F	A	F
3g	54.66	67.93	57.35	68.03	59.17	68.09	**60.43**	**68.91**	57.52	66.14	56.64	65.12
Baseline	50%		50%		50%		50%		50%		50%	

Table 2: UNIQUE_EXP Results

GRAMS+UNIQUE3+AUTHOR.

As Table 3 shows, when the HAS_POLAR feature is applied with the BASE TRIGRAMS configuration, the best *A* (i.e., 64.74%) is acquired using GI_LEX and the best *F* (i.e., 70.05%) is acquired when applying SentiWN_LEX. This is an improvement of 14.74% *A* and 20.05% *F* over BASE TRIGRAMS and 10.08% *A* and 2.19% *F*over the majority class baseline. As for the BASE TRIGRAMS+UNIQUE3 configuration, 64.21% *A* (with GI_LEX) and 69.55% *F* (with YT_LEX) are achieved. Although this is an improvement over the baseline, a slight degradation of performance (i.e., 0.53% *A* and 0.50% *F*) occurs as compared to the best results achieved with BASE TRIGRAMS.

Regarding the BASE TRIGRAMS+AUTHOR configuration, the best results of 95.51% *A* and 95.60% *F* are achieved using the YT_LEX. This is 45.51% *A* and 45.56% *A* improvement over the baseline. As Table 3 also shows, applying this configuration also improves over both the BASE TRIGRAMS and the BASE TRIGRAMS+UNIQUE3 configurations. The TRIGRAMS+UNIQUE3+AUTHOR achieves 94.78% *A* and 94.80% *F* with YT_LEX applied, which is a significant improvement over the baseline and a slight improvement (i.e., 0.07% over the *F* of the BASE TRIGRAMS).

From Table 3, it can be concluded that the best results are acquired using the BASE TRIGRAMS+AUTHOR configuration when the YT_LEX is employed. This shows that our manually-created YT_LEX outperforms the number of popular lexicons we test. We deduce that our lexicon is best suited to the health domain.

7 Related Work

A number of datasets have been labeled for SSA. Most relevant to us is work on the news genre. (Wiebe et al., 2005) label a news corpus at the word and phrase level, but neither label data for domain nor use the *Author* and *News source* we introduce here. (Balahur et al., 2009) label quotations from the news involving one person mentioning another entity and maintain that quotations typically contain more sentiment expressions than other parts of news articles. Our work is different from that of (Balahur et al., 2009) in that we label all sentences regardless whether they include quotations or not.

Many subjectivity tagging systems have also been proposed. For example, Wiebe et al. (Wiebe et al., 1999) attempt to classify news data for subjectivity, at the sentence level. useing POS features and lexical features. They report 72.17% accuracy, which is more than 20% points higher than a baseline accu-

	BASE TRIGRAMS		**+UNIQUE3**		**+AUTHOR**		**+UNIQUE+AUTHOR**	
	A	F	A	F	A	F	A	F
-HAS_POLAR	54.66	67.93	60.43	68.91	95.50	95.51	94.78	94.73
+HAS_POLAR (**GI_LEX**)	**64.74**	62.99	**64.21**	66.74	92.27	92.81	91.72	92.20
+HAS_POLAR (**YT_LEX**)	54.66	67.93	61.14	**69.55**	**95.51**	**95.60**	**94.78**	**94.80**
+HAS_POLAR (**HW_LEX**)	58.25	68.51	60.06	67.55	94.43	94.60	94.05	94.11
+HAS_POLAR (**SentiWN_LEX**)	64.73	**70.05**	64.02	67.30	93.34	93.74	92.08	92.43
+HAS_POLAR (**SentiWN_Strong_LEX**)	62.23	63.96	61.49	65.65	93.88	94.16	92.44	92.67
Baseline	50%		50%		50%		50%	

Table 3: POLAR_EXP Results

racy obtained by always choosing the majority class. Bruce & Wiebe (Bruce and Wiebe, 1999) performed a statistical analysis of the assigned classifications in the corpus reported in (Wiebe et al., 1999). The analysis showed that adjectives are statistically significantly and positively correlated with subjective sentences in the corpus.

8 Conclusion

In this paper, we present a corpus of professional articles annotated at the paragraph level for subjectivity and sentiment, as well as topic. We motivate SSA for professional news articles and summarize our annotation scheme. Our approach is unique in that we label the data with topics inspired by the Obama administration as part of its popularization of the OHRP. In addition, we present a subjectivity tagging system that exploits this data, making use of novel and cheap meta-data features (i.e., SOURCE and AUTHOR) that significantly boost system performance. Further, we introduce a wide-coverage polarity lexicon that performs better on the health-domain data as represented by our data set than a number of other popular lexicons. Our system performs very successfuly on the task, with 95.51% accuracy and 95.60% F-measure, beating a 50.00% baseline.

References

S. Baccianella, A. Esuli, and F. Sebastiani. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Seventh conference on International Language Resources and Evaluation, Malta. Retrieved May*, volume 25, page 2010.

A. Balahur and R. Steinberger. 2009. Rethinking Sentiment Analysis in the News: from Theory to Practice and back. *Proceeding of WOMSA*.

A. Balahur, R. Steinberger, E. van der Goot, B. Pouliquen, and M. Kabadjov. 2009. Opinion mining on newspaper quotations. In *2009 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, pages 523–526. IEEE.

A. Banfield. 1982. *Unspeakable Sentences: Narration and Representation in the Language of Fiction*. Routledge Kegan Paul, Boston.

R. Bruce and J. Wiebe. 1999. Recognizing subjectivity. a case study of manual tagging. *Natural Language Engineering*, 5(2).

M.T. Diab, L. Levin, T. Mitamura, O. Rambow, V. Prabhakaran, and W. Guo. 2009. Committed belief annotation and tagging. In *Proceedings of the Third Linguistic Annotation Workshop*, pages 68–73. Association for Computational Linguistics.

V. Hatzivassiloglou and J. Wiebe. 2000. Effects of adjective orientation and gradability on sentence subjectivity. In *International Conference on Computational Linguistics, (COLING-2000)*.

T. Joachims. 2008. Svmlight: Support vector machine. http://svmlight.joachims.org/, Cornell University, 2008.

J. Wiebe, R. Bruce, and T. O'Hara. 1999. Development and use of a gold standard data set for subjectivity classifications. In *Proc. 37th Annual Meeting of the Assoc. for Computational Linguistics (ACL-99)*, pages 246–253, University of Maryland: ACL.

J. Wiebe, T. Wilson, R. Bruce, M. Bell, and M. Martin. 2004. Learning subjective language. *Computational linguistics*, 30(3):277–308.

J. Wiebe, T. Wilson, and C. Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2):165–210.

J. Wiebe. 1994. Tracking point of view in narrative. *Computional Linguistics*, 20(2):233–287.