SENTIMENT RETRIEVAL ON WEB REVIEWS USING SPONTANEOUS NATURAL SPEECH

Jose Costa Pereira[1,2], *Jordi Luque*[1] *and Xavier Anguera* [1]

[1]Telefonica Research, Barcelona, Spain
[2]Department of Electrical and Computer Engineering, University of California San Diego, USA

josecp@ucsd.edu, jls@tid.es, xanguera@tid.es

ABSTRACT

This paper addresses the problem of document retrieval based on sentiment polarity criteria. A query based on *natural spontaneous speech*, expressing an opinion about a certain topic, is used to search a repository of documents containing favorable or unfavorable opinions. The goal is to retrieve documents whose opinions more closely resemble the one in the query. A semantic system based on the speech transcripts is augmented with information from full-length text articles. Posterior probabilities extracted from the article are used to *regularize* their transcription counterparts. This paper makes three important contributions. First, we introduce a framework for polarity analysis of sentiments that can accommodate combinations of different modalities, while maintaining the flexibility of unimodal systems, i.e. capable of dealing with the absence of any modality. Second, we show that it is possible to improve average precision on *speech transcriptions' sentiment retrieval* by means of regularization. Third, we demonstrate the strength and generalization of our approach by training regularizers on one dataset, while performing sentiment retrieval experiments, with substantial gains, on a collection of YouTube clips.

Index Terms— Sentiment analysis, spontaneous speech reviews, polarity, subjectivity, information retrieval.

1. INTRODUCTION

The advent of social media acted as an immense highway for a continuous flow of opinionated information accessible to all. Every minute people share personal feelings on the web; e.g. reviews on recently acquired items, gastronomical experiences, emotional events in their lives, religious beliefs, or simple matters of opinion on everyday issues. Platforms like Twitter, Youtube, Facebook or Google+ are on the epicenter of these sentimental waves of information[1]. The availability of this user-centric data allied to commercial interests has spurred research efforts in many directions. For example, the fashion industry is interested in knowing what is trendy for next season, governmental agencies need to detect potential security threats, techonology companies want to measure the acceptance of their products and detect market-specific issues with implications on their sales strategy. All of these share an underlying need to "feel the pulse" of an audience. This is broadly known as sentiment analysis or opinion mining and is usually split into two smaller tasks: subjectivity [1] and polarity [2]. While subjectivity measures *how much sentiment* information a user opinion contains, polarity focuses on analyzing whether the opinion is *positive or negative*. In this work we focus on the latter.

Previous work on sentiment analysis has been predominantly unimodal, mostly in the text domain [3, 4, 5]. While traditional natural language processing (NLP) tasks (e.g. text classification) often rely on topic models, subjectivity and polarity detection typically use earlier NLP models. In [1] several examples on sentiment analysis using bag-of-words (BoW) are shown, while a topic model approach is used in [6]. Sentiment analysis is also considered a more challenging task when compared to text classification, and often requires domain-specific data. This has lead to the development of many sentiment lexicons [7, 8, 9] and large annotated datasets [1, 10, 11]. A challenging area that, in our opinion, has not received enough attention, is how to incorporate multiple sources of information for sentiment analysis. To the best of our knowledge only a couple of studies have been made along that direction [12, 13]. In particular, an important area of research, is the task of automatic sentiment extraction from natural audio streams containing spontaneous speech. In the present work, we aim to address this issue by performing sentiment analysis on speech transcriptions extracted from video reviews. With regards to previous multimodal sentiment analysis work, we would emphasize the following key distinctions: in our work different data sources are mapped onto a common space where sentiment is extracted, and although our retrieval framework can incorporate multimodal data, we note that it is robust enough to work in the absence of any modality.

In this paper, we propose a feature regularizer suitable for

At the time of this work, J. Costa Pereira was visiting Telefonica Research

[1]YouTube alone claims that more than 100 hours of video are being uploaded every minute, http://www.youtube.com/yt/press/statistics.html

sentiment polarity analysis through modal expansion. A set of reviews for small electronic appliances was collected, containing a small video clip and a full text article on each product. A linear operator is then learned to minimize the average similarity of data across the two modalities. This acts as a feature regularizer for samples belonging to the noisier of the modalities. Among other results, we show in section 3.1, how the regularization of transcriptions extracted from spontaneous natural speech is able to produce over 5% gains on retrieval accuracy of *unseen* YouTube transcriptions.

2. SENTIMENT ANALYSIS

In the present work we focus on automatic sentiment analysis in a scenario where multiple sources of information are available during the training stage (but not for testing). This does in fact exist in many practical applications; e.g. specialized review sites, social media, product review blogs. One example is the CNET database [14], a repository of product reviews with information available in two formats: a video with someone reviewing a consumer electronics product, and a full text article, not necessarily written by the same person, where a more thorough evaluation of the product is presented. We are interested in consolidating the prevailing sentiment polarity about a certain product expressed through these multiple sources (e.g. video, audio, textual).

2.1. Sentiment space

The simplest approach to represent multi-modal data consists in concatenation of features from different sources. This however raises questions on how to deal with missing modalities. Another approach is to map data onto a common space, where heterogeneous modalities are represented with the same features [15]. This space, *made by design*, is tailored to the task at hand and is the approach we follow in this work. For opinion mining we leverage heterogenous sources of information into a *sentiment space*. This is done by building classifiers for each modality that map a low-level feature space, $\mathcal{R}$, onto a vector of class posterior probabilities in the sentiment space, $\mathcal{S}$; where dimensions represent the polarity expressing either positive or negative sentiments.

Figure 1 summarizes how to obtain the sentiment polarity representation for both modalities: text and video reviews. In the case of a text snippet, the bag-of-words model can be obtained directly from low-level features (e.g. words, bigrams). For the video review, an automated speech recognition (ASR) system is first used to estimate the corresponding transcriptions. These are usually noisy text features, and are used as low-level representation for this modality. The existing gap between the accuracy of the two modalities allows for significant gains on sentiment retrieval by means of regularization. A framework that is introduced in the following section.

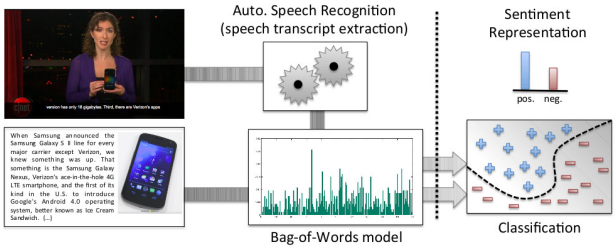
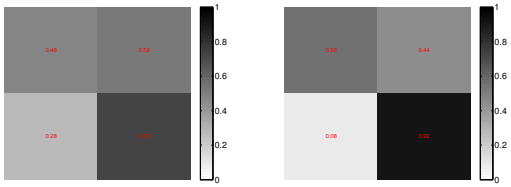


Fig. 1: To obtain the sentiment polarity representation low-level features are first extracted to obtain a bag-of-words. A text classifier maps this to $\mathcal{S}$, the space of sentiment polarities.

2.2. Polarity regularization

The principle of modal expansion relies on the fact that one (auxiliar) modality may contain more information about the sentiment than the data in which we are interested [16]. In the present study, the data extracted by an ASR is not as informative in predicting the polarity of the prevalent sentiment as a full text article. This can be noted by inspection of the confusion matrices on the individual modalities, as shown on Figure 2. We use this principle to find a transformation that *morphs* posterior probabilities obtained from the less accurate modality (Fig. 2a) to the ones obtained through the more accurate data (Fig. 2b); e.g. speech transcriptions and full text article, respectively.



(a) Transcriptions. (b) Articles.

Fig. 2: Confusion matrices on unimodal classification of the speech transcriptions and text articles from the CNET dataset.

More formally, given a set of observations for product reviews in two modalities $(\mathbf{x}_s^i, \mathbf{x}_t^i)$ with $i = 1 \dots n$, where $\mathbf{x}_s^i$ is the i^{th} sample of the speech transcription, and $\mathbf{x}_t^i$ the corresponding observation of the text article, we first obtain their representation onto the space of sentiment class posteriors through suitable unimodal classifiers,

$$\boldsymbol{\pi}_s^i \leftarrow \Gamma(\mathbf{x}_s^i) \qquad \text{and} \qquad \boldsymbol{\pi}_t^i \leftarrow \Psi(\mathbf{x}_t^i)$$

where $\Gamma(.)$, $\Psi(.)$ stand for the classifier function for speech transcriptions, text articles, respectively; and $\boldsymbol{\pi}_s^i$, $\boldsymbol{\pi}_t^i$, is the i^{th} 2-dimensional posterior vector obtained as a result of applying the classifier to the low-level representation $\mathbf{x}_s^i$, $\mathbf{x}_t^i$, respectively. Then, we partition the observations in two groups, positive $(+)$ and negative $(-)$ reviews. Using reviews that exhibit similar polarity, we learn two transformations Φ^j, $j \in$

$\{+,-\}$ on the sentiment polarity space ($\mathcal{S}$)

$$\begin{array}{rccccl} \Phi^j: & \mathcal{S} & \rightarrow & \mathcal{S} & & j \in \{+,-\} \\ & \boldsymbol{\pi}_s & \rightarrow & \boldsymbol{\pi}_t & & \end{array}$$

A solution for each Φ^j can be found in the least squares sense by solving

$$\min \parallel \Phi^j(\boldsymbol{\pi}_s^j) - \boldsymbol{\pi}_t^j \parallel_2^2 \quad (1)$$

where positive sample pairs $(\boldsymbol{\pi}_s^+, \boldsymbol{\pi}_t^+)$ are used to learn Φ^+, and negative pairs are used to learn Φ^-. Restricting Φ^j to the class of linear operators, it can be shown that equation (1) can be mapped to a quadratic programming problem with affine constraints. These constraints enforce the *range space* of Φ^j to lie on the simplex, making any transformed point a valid probability vector. Furthermore, under certain mild conditions, the feasible set and the optimization problem are convex and a global minimum can be obtained using the approach described in [16].

After this procedure two matrices are obtained; Φ^+ and Φ^- regularize towards positive and negative sentiment classes by linear projection of the original probability posterior vectors. Now we focus on the problem of applying these projections to *any speech transcript sample* ($\boldsymbol{\pi}_s$) given on the sentiment polarity space.

2.3. Weighting the transformations

In this work we consider the usual two classes for polarity: positive (+) or negative (−) sentiments. The regularization procedure introduced in the previous section provides the tools to obtain a suitable transformation for each class of polarity, i.e. Φ^+ (positive) and Φ^- (negative). To obtain the regularized features for a given speech transcript, $\boldsymbol{\pi}_s$, in general, we don't know which transformation to apply. Therefore we use a convex combination of the possible transformations, yielding the regularized features:

$$\hat{\boldsymbol{\pi}}_s = \sum_{j \in \{+,-\}} w^j \, \Phi^j(\boldsymbol{\pi}_s) \quad (2)$$

In equation (2) a vector of weights, $\mathbf{w}$, is used to interpolate the regularization transformations. To determine $\mathbf{w}$ we consider two possibilities: in a complete scenario, the sample point $\mathbf{x}_s$ has an associated text review, $\mathbf{x}_t$, in which case we set $\mathbf{w} \leftarrow \boldsymbol{\pi}_t = \Gamma(\mathbf{x}_t)$; if there is no associated text review we perform a cross-modal search [17] operation, using $\mathbf{x}_s$, to find the most suitable $\mathbf{x}_t$, which will then yield $\mathbf{w}$ in a way similar to the complete scenario. In such operation, a similarity measure is computed in a pairwise comparison of the $\boldsymbol{\pi}_s$ point to each available text article reviews $\mathcal{T} = \{\boldsymbol{\pi}_t^1, \boldsymbol{\pi}_t^2, \ldots, \boldsymbol{\pi}_t^n\}$. Then, $\mathbf{w}$ is fixed as follows:

$$\begin{array}{rcll} \mathbf{w} & = & \boldsymbol{\pi}_{j*} & \\ j* & = & \arg\max_j S(\boldsymbol{\pi}_t^j, \boldsymbol{\pi}_s) & \forall\, j = 1 \ldots n. \end{array}$$

In this work $S(a,b)$ is the *Kulback-Leibler* divergence, but any suitable similarity measure between a and b can be adopted.

3. EXPERIMENTAL SETUP AND RESULTS

In this section experiments are conducted to assess whether an additional source of information can be used to improve results on the task of retrieving reviews that express similar sentiment to a given query. All experiments refer to a query-by-example (QBE) retrieval scenario in the sentiment space. Speech transcripts are used as queries. Different scenarios are considered for the retrieval set.

Datasets: The CNET [14] dataset is a public collection of consumer electronic product reviews available in two modalities: 1) a full text article about the product, written by one of CNET's editors, and 2) a video where CNET people describe their evaluation on that product. In particular, we restrict to the set of product reviews whose videos contained closed captions representing the speech transcriptions. In both modalities – speech transcriptions and text reviews – pros and cons of each product are detailed and an overall editor rating of the product is made. The rating ranges from 1 to 5 stars[2]. Although the reviews tend to be somewhat lenient, we have set the threshold between a good and bad review at the mean point 2.5, i.e. any product rated 2.5 or below would be considered "negative" and the rest are considered "positive" reviews. Our final CNET collection contains 50 reviews on both modalities, 25 from each "sentiment" as summarized in Table 1.

source	words	sentences	samples
speech transcripts	475	22	50
text articles	2330	85	50

Table 1: CNET document statistics: average number words and sentences per document, broken down per modality.

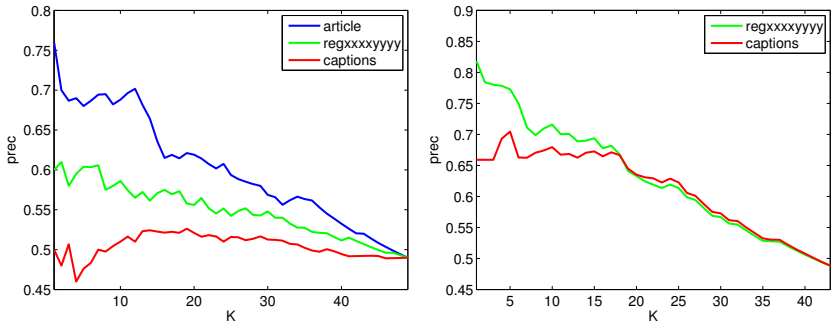
The other dataset used in this work is a collection of YouTube videos where people express opinions about a wide variety of topics. This dataset results from the concatenation of positive and negative reviews available in [13, 18]. From the 28 labelled videos of [18], we used 20 videos that were still available online at the time of the experiments. From [13] we use a total of 24 videos, 12 videos from each class. This yields a total of 44 sentiment videos for the YouTube experiments. Please refer to the original papers [13, 18] for more details on each set of videos.

Representation: Audio transcriptions are available in the CNET video reviews in the form of closed captions. For the

[2] 1 and 5 star rated products are hard to find. Our final collection contains products ranging 2 to 4.5 stars.

	Dataset	
	CNET	Youtube
Article	0.678	n/a
Transc.$_{\text{reg}}$	0.624	0.726
Transcript	0.546	0.689
random	0.5	0.5

Table 2: Mean average precision (mAP) scores for retrieval on both datasets.



(a) CNET reviews (b) YouTube reviews

Fig. 3: Precision @K for both datasets (CNET and YouTube); i.e. ratio of correct samples on the top K retrieved results. Note that text article reviews are not available on the YouTube data.

YouTube dataset, the automatic captions feature is used to generate the speech transcripts[3]. Both text articles and speech transcripts are english written reviews. The popular BoW model is used to represent them at a low-level space. We follow the best practices in text representation for sentiment analysis [1] and consider both unigrams and bigrams, no stop-word list, no stemming nor lemmatization.

To obtain the representation on the sentiment space we train a Naïve Bayes classifier using the software package of [19]. The choice of the "best" model lies outside the scope of this work, nevertheless we note that more complex models can be used without implications in the current framework. A comparison between different models can be found in [1]. For increased robustness we use yet another dataset for training the BoW model. A random sample of $3,000$ observations from the Amazon reviews dataset [20]; $1,500$ samples with 5 stars reviews and the same amount for reviews rated with 1 star. We note that the same classifier is used throughout the experiments, regardless of the dataset being tested, i.e. CNET or YouTube. This shows a fair amount of generalization on our framework. The classifier yields the 2-dimensional vector, $\boldsymbol{\pi}$, that expresses the belief that a certain review contains positive (negative) sentiments.

3.1. Results

Speech transcriptions are used as queries on a QBE retrieval scenario. The first set of experiments uses CNET reviews [14], where three retrieval database scenarios are considered: (1) full text articles, (2) speech transcripts from the video clip, and (3) regularized transcripts. For the second set of experiments we use the YouTube video reviews collected by [18, 13]. In this dataset there is no corresponding text article associated with the videos. Hence, we consider retrieval experiments on speech transcripts data, and on regularized transcript features. Standard information retrieval metrics [21] are used to assess performance. Table 2 shows mean average precision (mAP) for each scenario. Summarizing retrieval precision in a single number. Figure 3 shows precision for all values of K on each experiment.

CNET: Regularization of the speech transcripts achieves a gain of 14% relative to the non-regularized transcripts, raising mAP from 0.55 to 0.62. "Transc.$_{\text{reg}}$" features are around 9% below the upper bound mAP achieved when retrieving on the full text articles; we note, however, that the full article reviews are seldom available in real life applications. From Figure 3a one sees that the overwhelming gain on precision comes from lower values of K. This favors a scenario where people only look at the first results of a search operation.

YouTube: Retrieval accuracy for the YouTube dataset is shown on the rightmost column of Table 2. Note that only one modality exists in this dataset. It is worth mentioning the higher mAP for transcriptions obtained for this dataset as compared to the previous dataset. This is likely due to the more lenient reviews made by CNET professionals. The same operators learned for CNET data are used here. Since this dataset only has speech transcript data, for the transformation weighting we use the cross-modal search strategy introduced in section 2.3. The mAP gains are relatively (5.5%) more modest when compared to the previous experiment, going from 0.69 to 0.73. The precision@K curve of Figure 3b exhibits the same desirable effect as before, i.e. higher gains for lower values of K.

4. CONCLUSIONS AND FUTURE WORK

A recent work on sentiment analysis has shown benefits in joining several sources of information [13], in line with these results we present a learning framework that transfers knowledge across modalities. Our approach was shown to improve similar-sentiment retrieval accuracy even in the absence of one modality. Benefits are achieved in the overall average precision, more specifically for lower values of retrieved documents. As noted by [22], ignoring word order – such as

[3]https://support.google.com/youtube/answer/3038280

in the BoW representiation – in the treatment of a semantic task is not plausible, and it may result in serious damage of sentiment analysis in hard cases of negation. In the future, it is planned the extension of this sentiment polarity analysis framework in order to accommodate more modalities, e.g. acoustic and visual related features, and to explore text articles at the sentence level with a coherent analysis transferred to the whole document.

5. REFERENCES

[1] Bo Pang and Lillian Lee, "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts," in *Proc. meeting ACL*. ACL, 2004, p. 271. 1, 4

[2] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis," in *Proc. Conf. Human Language Technology and EMNLP*. ACL, 2005, pp. 347–354. 1

[3] Minqing Hu and Bing Liu, "Mining and summarizing customer reviews," in *Proc. SIGKDD*. ACM, 2004, pp. 168–177. 1

[4] Krisztian Balog, Gilad Mishne, and Maarten de Rijke, "Why are they excited?: identifying and explaining spikes in blog mood levels," in *Proc. Conf. European Chapter ACL: Posters & Demos*. ACL, 2006, pp. 207–210. 1

[5] Namrata Godbole, Manja Srinivasaiah, and Steven Skiena, "Large-scale sentiment analysis for news and blogs.," in *Int. Conf. Weblogs and Social Media*, 2007, vol. 7. 1

[6] Jordan Boyd-Graber and Philip Resnik, "Holistic sentiment analysis across languages: Multilingual supervised latent dirichlet allocation," in *Proc. Conf. EMNLP*, 2010. 1

[7] Philip J Stone, Dexter C Dunphy, and Marshall S Smith, *The General Inquirer: A Computer Approach to Content Analysis.*, MIT press, 1966. 1

[8] Janyce Wiebe and Ellen Riloff, "Creating subjective and objective sentence classifiers from unannotated texts," in *Comp. Linguistics and Intelligent Text Processing*, pp. 486–497. Springer, 2005. 1

[9] Andrea Esuli and Fabrizio Sebastiani, "Sentiwordnet: A publicly available lexical resource for opinion mining," in *Proc. LREC*, 2006, vol. 6, pp. 417–422. 1

[10] Janyce Wiebe, Theresa Wilson, and Claire Cardie, "Annotating expressions of opinions and emotions in language," *Language Resources and Evaluation*, vol. 39, no. 2-3, pp. 165–210, 2005. 1

[11] Carlo Strapparava and Rada Mihalcea, "Semeval-2007 task 14: Affective text," in *Proc. Int. Workshop on Semantic Evaluations*. ACL, 2007, pp. 70–74. 1

[12] Stephan Raaijmakers, Khiet Truong, and Theresa Wilson, "Multimodal subjectivity analysis of multiparty conversation," in *Proc. Conf. EMNLP*. ACL, 2008, pp. 466–474. 1

[13] Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi, "Towards multimodal sentiment analysis: Harvesting opinions from the web," in *Proc. Int. Conf. Multimodal Interfaces*. ACM, 2011, pp. 169–176. 1, 3, 4

[14] CBS Interactive Inc., "Product reviews, CNET," 2013. 2, 3, 4

[15] Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Nikhil Rasiwasia, Gert Lanckriet, Roger Levy, and Nuno Vasconcelos, "On the role of correlation and abstraction in cross-modal multimedia retrieval," *Trans. PAMI*, 2013. 2

[16] J. Costa Pereira and N. Vasconcelos, "On the Regularization of Image Semantics by Modal Expansion," in *CVPR*, 2012, pp. 3093–3099. 2, 3

[17] Nikhil Rasiwasia, Jose Costa Pereira, Emanuelle Coviello, Gabriel Doyle, Gert Lanckriet, Roger Levy, and Nuno Vasconcelos, "A New Approach to Cross-Modal Multimedia Retrieval," in *Proc. Int. Conf. on Multimedia*. ACM, 2010, pp. 251–260. 3

[18] Lakshmish Kaushik, Abhijeet Sangwan, and John HL Hansen, "Sentiment extraction from natural audio streams," in *Proc. Int. Conf. ASSP*, 2013. 3, 4

[19] Jacob Perkins, "Nltk trainer," 2011. 4

[20] Nitin Jindal and Bing Liu, "Opinion spam and analysis," in *Proc. Int. Conf. Web Search Web Data Mining*. ACM, 2008, pp. 219–230. 4

[21] C.D. Manning, P. Raghavan, and H. Schütze, *An Introduction to Information Retrieval*, Cambridge University Press, 2008. 4

[22] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Chris Manning, Andrew Ng, and Chris Potts, "Recursive deep models for semantic compositionality over a sentiment treebank," in *Proc. Conf. EMNLP*, 2013. 4