Proc. of The Third Intl. Conf. on Advances in Computing, Electronics and Communication - ACEC 2015

ISBN: 978-1-63248-064-4 doi: 10.15224/ 978-1-63248-064-4-16

Sentiment Analysis: Publicly available datasets

Diana C. Cavalcanti

***Abstract*—Sentiment analysis includes computational techniques to understand opinions and sentiment in texts. Several studies have explored supervised and unsupervised methods for sentiment classification covering natural language processing techniques, information retrieval and lexical resources. Training and testing data are appropriate from documents since each review has already labeled. This paper describes labeled datasets publicly available for sentiment classification problem.**

***Keywords*— sentiment analysis, dataset, public, online**

I. Introduction

Sentiment Analysis (SA) explores the computational study of opinions, sentiments and emotions expressed in sources such as unstructured text. SA describes what must be extracted from sources of opinion, for example, online reviews, consumer opinions of forums, ecommerce, social networking, blogs or political reviews and how the results can be organized and presented to the user [18, 19, 20, 21].

Pang and Lee [22] comment "what others think" has always been an important piece of information for most people during the process of decision making. Sentiment analysis has been used for different practical applications such which recommender systems, summarization, political systems, ranking of products or services, marketing intelligence since many companies are interested in tracking your brand and market awareness [18, 21].

As the field of sentiment analysis involves classifying opinions in text into categories like "positive", "negative" or "neutral" from an appraiser in relation to a subject, object or person [23].

The task classification can be done in word level, sentence or document. For instance, sentiment analysis can be applied to words or phrases and then use this information to evaluate sentences or documents.

Several studies have explored supervised and unsupervised methods for classification of sentiment that have involved techniques of Natural Language Processing, Information Retrieval and Lexical Features, such techniques can be used together to address the issues in sentiment analysis, classification and summarization [18, 21, 22, 23, 24].

Diana C. Cavalcanti
Centro de Informática/Universidade Federal de Pernambuco, Brazil

Improve accuracy of sentiment analysis methods requires training and testing data to assess their performances. The datasets applied in SA are a relevant item in this field. Recently several evaluation datasets from blogs and product reviews have been made publicly available.

Training and testing data are appropriate from these documents since each review has already labeled. The rating to SA datasets has been annotated with different sentiment labels including: "Negative", "Neutral", "Positive" or numerical rating -1, 0, +1 to represents negative, neutral, positive respectively. Other typical label is range of polarity, for example 1-5 stars for movies or products [18, 21].

In this paper we present relevant datasets publicly available online to the problem of sentiment analysis. Our goal is to provide an overview of the existing community datasets and their properties to study and research. We have highlighted some features: (i) publicly available to the research community, (ii) format available, (iii) sentiment classification (example: numerical binary, ternary, stars), (iv) size of dataset. All datasets cited are in English language. In the section 4 are cited two collections in different language. The table 5 describes online addrress to each dataset cited.

The datasets are made available for non-commercial and research purposes only. All rights, including copyright, in the content of the original abstracts are owned by the original authors.

The remaining of this paper is organized as follows. Section II presents datasets crawled from products reviews, followed Section III which describes datasets about movie reviews. Section IV provides datasets extracted from Twitter.com. Finally, Section V concludes the paper.

II. Product Reviews

Internet users have become not only a consumer of a product, but also a generator of Web content.

People want to know what are the feelings that other person expressed in relation to the consumption of certain products and services to assist in their consumer decisions [19, 20]. In this section are cited dataset that include labeled reviews of products like cellphone, electronics, cars and games.

A. Multi-Domain Sentiment Dataset

The database The Multi-Domain Sentiment version 2.0 contains product reviews extracted from Amazon.com. The base has reviews distributed in 25 different fields (eg, camera and photo, automobiles, mobile devices and services, musical instruments, clothing). For each domain, the comments are separated into three files: positive (positive.review), negative (negative.review) and unclassified (unlabeled.review).

Proc. of The Third Intl. Conf. on Advances in Computing, Electronics and Communication - ACEC 2015

ISBN: 978-1-63248-064-4 doi: 10.15224/ 978-1-63248-064-4-16

The base has a total of 142.253 million reviews, of which 38.458 are labeled documents distributed on the positive and negative classes. Each review has a star rating (1-5 stars). Negative comments have 1 or 2 star rating, while positive comments have 4 or 5 star rating. The comments are organized in a pseudo (Extensible Markup Language) XML schema (Blitzer et al., 2007). Figure 1 illustrates an example of review.

```
<review>
<unique_id>
B000E18BTM:battery_consuming_product:s._nilawar
</unique_id>
<asin>B000E18BTM</asin>
<product_name>
Panasonic Lumix DMC-LZ3S 5MP Digital Camera with 6x Image Stabilized Zoom: Camera & Photo
</product_name>
<product_type>camera & photo</product_type>
<helpful>1 of 5</helpful>
<rating>2.0</rating>
<title>Battery consuming product</title>
<date>January 11, 2007</date>
<reviewer>S. Nilawar</reviewer>
<reviewer_location></reviewer_location>
<review_text>I am not satisfied by this product. The calrity is not good. Also this is the most battery consuming device I have ever seen
</review_text>
</review>
```

Figure 1. Example of review of The Multi-Domain Sentiment

B. Opinion Mining/Sentiment Mining Datasets

Sentiment Mining datasets contain a total of 6.034 reviews from different sources including product reviews and other domains as described:

- *Digital camera* with 498 reviews (Digital camera reviews from Amazon.com);
- Summer camp with 804 reviews (Summer camp reviews from CampRatingz.com);
- *Physician* with 1.478 reviews (Reviews of physicians from RateMDs.com);
- *Pharmaceutical* drug with 802 reviews (Reviews of pharmaceutical drugs from DrugRatingz.com),
- *Laptop* with 176 reviews, (Laptop reviews from Amazon.com);
- *Lawyer* with 220 reviews (Reviews of lawyers from LawyerRatingz.com);
- *Music (DC)* with 582 reviews (Musical CD reviews from Amazon.com),
- *Radio show* with 1004 reviews (Reviews of radio shows from RadioRatingz.com)
- *TV show* with 470 reviews (Television show reviews from TVRatingz.com)

Each review has a numerical rating -1.0 to negative reviews and 1.0 to positive reviews and is organized in .txt files. Each domain contain 50% of positive reviews and 50% of negative reviews [6]. The table 1 presents one example of review to this dataset.

TABLE I. REVIEW OF THE OPINION MINING/SENTIMENT MINING DATASETS

Review	Rating
This camera is amazing! The 7.1 megapixels make extremely clear pictures and the size is great for a point-and-shoot camera. It is working out great for business and personal!	1.0

C. Web data: Amazon reviews

This corpus has 34.686.770 reviews from Amazon.com distributed in 34 categories (eg. software, musical instruments, eletronics, beauty, jewelry, shoes, pet supplies, watches, etc.). The reviews have a review/scores 1-5 rating. And each domain is organized in .txt file.

Also available are other databases from Amazon.com in http://snap.stanford.edu/data/#reviews including reviews about beers, fine foods and wine. And in http://jmcauley.ucsd.edu/data/amazon/ provides dataset which corrects the above duplication issues, all datasets in similar format as described in Figure 2.

```
product/productId: B000GX8THM
product/title: Invicta Women's 8940 Pro Diver Collection Watch
product/price: 42.49
review/userId: A3ORKADM4TUKQK
review/profileName: Plantronics m155
review/helpfulness: 0/0
review/score: 4.0
review/time: 1353801600
review/summary: Good Watch
review/text: I love this product and I am glad I purchased it.I will always buy this product and recommend it to anyone.
```

Figure 2. Web data: Amazon reviews: sample review

D. Finegrained Sentiment Data Set, Release 1

This dataset correspond a set of reviews manually annotated at the sentence level. Each sentence is classified in five categories: POS (sentences as positive), NEG (sentences as negative), NEU (sentences that express sentiment, but are neither positive nor negative), MIX (sentences that express both positive and negative sentiment), and NR (sentences not relevant).

The data contain 294 product reviews and 3.836 sentences of five products: books, dvds, electronics, music, videogames. The data is organized in one unique .txt file in format presented in table 2 [10, 11].

ISBN: 978-1-63248-064-4 doi: 10.15224/ 978-1-63248-064-4-16

TABLE II. THE FINEGRAINED SENTIMENT DATA SET : REVIEW

Rating	Sentence
mix	The action scenes were cool, the story was alright, but all in all i was not impressed.
neg	However, It is a little disappointed about dis movie.
nr	She gives him three cars totalling more than 750,000 dollars.

E. Opinion-claims

Opinion-claims dataset is a collection of customer reviews of five products: Canon G3, Nikon coolpix 4300, Nokia 6610, Creative Labs Nomad Jukebox Zen Xtra 40GB, Apex AD2600 Progressive-scan DVD player.

The data contains total of 1.252 reviews distributed in 383 negative and 869 positives reviews available in .csv file.

F. Micropinion Generation Dataset

This dataset contain reviews extracted from CNET.com. The reviews are about products from various categories like tv, cell phones, gps. Classified reviews are labeled in "pros" or "cons" and are available in .raw file [11].

G. LARA Review Dataset

Three datasets are presented in three categories. Each review is rated with ranges from 0 to 5 stars [25, 26]:

a. TripAdvisor Data Set

This dataset contains 235.793 hotel reviews crawled from TripAdvisor. The data can be download in .json or text file.

b. Amazon MP3 Data Set

The data contain 55.740 reviews about mp3 players extracted from Amazon.com. The data can be download in text file.

c. Six Categories of Amazon Product Reviews

The data contain Amazon product reviews of six categories: camera, mobile phone, TV, laptop, tablet and video surveillance system. The data can be download in .json file.

H. Customer Review

The data contain annotated customer reviews about different products, e.g..: digital camera, cellular, phone, mp3 player, dvd player, norton. Each category is available in separated text file and it is labeled with numeric score [+n] or [-n] where n is the opinion strength, 3 is strongest, and 1 is weakest.

III. Movies Reviews

In this section are presented collections of movie reviews. Each collection contains personal comments about films and aspects related.

A. Large Movie Review Dataset

This dataset provides 50.000 reviews of movie reviews associated with binary sentiment polarity labels. The data is separated in train e test sets. Each file is available in .txt file where each one is associated with 25.000 positive reviews or negative reviews [13].

B. Web data: Amazon movie reviews

The Amazon movie reviews contain 7.911.684 reviews from Amazon.com rated 1-5 scores. And each domain is organized in .txt file with same format presented in figure 2. [14].

C. Sentiment Analysis on Movie Reviews

The Sentiment Analysis on Movie Reviews was extracted from the Rotten Tomatoes dataset originally collected by Pang and Lee [15]. The data contain 156.060 phrases labeled as follow: 0 to negative (7.072 phrases), 1 to somewhat negative (27.273 phrases), 2 to neutral (79.582 phrases), 3 to somewhat positive (32.927 phrases), 4 to positive (9.206 phrases). Table 3 presents one example of review in dataset Sentiment Analysis on Movie Reviews.

TABLE III. THE SENTIMENT ANALYSIS ON MOVIE REVIEWS: EXAMPLE OF REVIEW

Phrase Id	Sentence Id	Phrase	Sentiment
43622	2112	Good performances and a realistic, non-exploitive approach make Paid in Full worth seeing.	4

D. Movie Review Data

Movie Review Data provides multiple datasets labeled in sentence or reviews level. The data provide data classified related at sentiment (positive or negative), subjectivity rating (stars) or subjectivity status (subjective or objective).

IV. Twitter

In this section collections including personal comments extracted from Twiter.com are presented.

A. Sentiment140

Sentiment140 is a collection of tweets extracted from Twiter.com classified in four categories: 0 = negative, 2 = neutral and 4 = positive. Each tweet was labeled based in emotions, for example tweets with emotion :) were classifieds positive and tweets with emotion :(were classifieds negative [16].

ISBN: 978-1-63248-064-4 doi: 10.15224/ 978-1-63248-064-4-16

TABLE IV. EXAMPLE OF REVIEW FROM THE SENTIMENT140

Polarity	Id	Date	Query	User	Text
4	1008	Mon May 11 03:29:01 UTC 2009	obama	lingbellbell	Obama is quite a good comedian! check out his dinner speech on CNN :) very funny jokes.

B. Twitter Sentiment Analysis Dataset

The dataset includes classified data of two sources: University of Michigan Sentiment Analysis competition on line (https://inclass.kaggle.com/c/si650winter11) and Kaggle Twitter Sentiment Corpus by Niek Sanders (http://www.sananalytics.com/lab/twitter-sentiment/). Were tagged manually 1.578.627 tweets as 1 for positive sentiment and 0 for negative sentiment.

C. Twitter Data set for Arabic Sentiment Analysis Data Set

This source provides labeled tweets in Arabic language date. There are 1000 positive and 1000 negative tweets separated in .txt files.

D. TASS 2013 Corpus

Two datasets in Spanish language are available. First one, the general corpus contains over 68.000 tweets about 150 personalities and celebrities of the world, politics, economy, communication, mass media and culture. Each message is classified in one of six level defined: strong positive (P+), positive (P), neutral (NEU), negative (N), strong negative (N+) and no sentiment tag (NONE).

Second one is the Politics corpus that contains 2.500 tweets about the electoral campaign of the 2011 General Elections in Spain. The messages are labeled in positive (P), neutral (NEU), negative (N) or no sentiment tag (NONE). The data are available in .xml file.

v. Conclusion

Sentiment Analysis problem has been a research interest for recent years and also involves practical applications in various fields. Research involved to automate tasks of classification of sentiment require efficiency in the preparation of the data to be trained. Such methods that require large formation of labeled data takes time and can be expensive to get labeled data, moreover, reduce the size of the training data may result in reduced performance of the classifier. This paper described labeled datasets publicly available for sentiment classification tasks.

References

[1] S. Park, W. Lee and I.C. Moon, "Efficient extraction of domain specific sentiment lexicon with active learning," Pattern Recognition Letters, 56(4), pp.38–44.

[2] P. Bharathi and PCD. Kalaivaani, "Incremental Learning on Sentiment Analysis Using Weakly Supervised Learning Techniques," IJESIT, vol. 3, Issue 2, 2014.

[3] S. Gao and H. Li, "A cross-domain adaptation method for sentiment classification using probabilistic latent analysis," CIKM, 2011.

[4] M. Marchand and R. Besançon, O. Mesnard and A. Vilnat, "Domain Adaptation for Opinion Mining: A Study of Multipolarity Words," JLCL, vol. 29, n. 1, pp. 17-31, 2014.

[5] J. Blitzer, M. Dredze and F. Pereira, "Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification," ACL, 2007.

[6] M. Whitehead and L. Yaeger, "Building a General Purpose Cross-Domain Sentiment Mining Model,Computer Science and Information Engineering," vol.4, pp.472-476, 2009.

[7] S. Arora, M. Joshi and CP Rose, "A Multi-dimensional Annotation Scheme for Opinion Mining from Unstructured Data," GSC, doi=10.1.1.208.3192, 2008.

[8] M. Hu and B. Liu, "Mining and summarizing customer reviews," ACM SIGKDD, KDD-04, 2004.

[9] M.g Hu and B. Liu, "Mining Opinion Features in Customer Reviews," AAAI, 2004.

[10] O. Täckström and R. McDonald, " Discovering fine-grained sentiment with latent variable structured prediction models," ECIR, 2011.

[11] K. A. Ganesan, C. X. Zhai, and E. Viegas, "Micropinion Generation: An Unsupervised Approach To Generating Ultra-Concise Summaries Of Opinions," 21st International Conference on World Wide Web, 2012.

[12] S. Arora, M. Joshi and C. Rose, "Identifying Types of Claims in Online Customer Reviews," NAACL, 2009.

[13] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, vol. 1, pp. 142-150, 2011.

[14] J. McAuley and J. Leskovec, "From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews," WWW, 2013.

[15] B. Pang and L. Lee, "Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales," ACL, pp. 115–124, 2005.

[16] A. Go, R. Bhayani and L. Huang, "Twitter Sentiment Classification using Distant Supervision," echnical report, Stanford, 2009.

[17] N. A Abdulla., N. A Mahyoub., M. Shehab and, M.Al-Ayyoub, "Arabic Sentiment Analysis: Corpus-based and Lexicon- based," AEECT, 2013.

[18] B. Liu, "Sentiment Analysis Mining Opinions, Sentiments, and Emotions," Cambridge University Press; 1 ed., 2015.

[19] E. Boiy, P. Hens, K. Deschacht, M. F. Moens, "Automatic sentiment analysis in on-line text," 11th International Conference on Electronic Publishing, ELPUB, pp. 349-360, 2007.

[20] E. Boiy, M. Moens, "A machine learning approach to sentiment analysis in multilingual Web texts," Information Retrieval Journal, vol. 12, n.5, pp. 526-558, 2008.

[21] A survey on opinion mining and sentiment analysis: Tasks, approaches 4 and applications, 2015.

[22] Pang, B.; Lee, L. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, v.2, n.1-2, pp. 1-135, 2008.

[23] S. Vohra and J. Teraiya, Applications and Challenges for Sentiment Analysis : A Survey, International Journal of Engineering Research & Technology, vol. 2, Issue 2, 2013.

[24] S. Padmaja and S. S. Fatima, "Opinion Mining and Sentiment Analysis – An Assessment of Peoples' Belief: A Survey," IJASUC, vol. 4, n. 1, 2013.

[25] H. Wang, Y.Lu and C. Zhai, "Latent Aspect Rating Analysis without Aspect Keyword Supervision,".17th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 618-626, KDD, 2011.

[26] H. Wang, Y. L. and C. Zhai, "Latent Aspect Rating Analysis on Review Text Data: A Rating Regression Approach," 16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining , KDD, pp.783-792, 2010.

ISBN: 978-1-63248-064-4 doi: 10.15224/ 978-1-63248-064-4-16

TABLE V. WEB ADDRESS OF THE DATASETS

DataSet	Available in:
Product Reviews	
Multi-Domain Sentiment Dataset	http://www.cs.jhu.edu/~mdredze/datasets/sentiment/.
Opinion Mining/Sentiment Mining Datasets	https://personalwebs.coloradocollege.edu/~mwhitehead/html/opinion_mining.html
Web data: Amazon reviews	http://snap.stanford.edu/data/web-Amazon-links.html
Finegrained Sentiment Data Set, Release 1	https://github.com/oscartackstrom/sentence-sentiment-data
Opinion-claims	http://www.cs.cmu.edu/~shilpaa/datasets/opinion-claims/
Micropinion Generation Dataset	http://kavita-ganesan.com/content/micropinion-generation-dataset
LARA Review Dataset	http://sifaka.cs.uiuc.edu/~wang296/Data/index.html
Customer Review	http://www.cs.uic.edu/~liub/FBS/CustomerReviewData.zip http://www.cs.uic.edu/~liub/FBS/Reviews-9-products.rar
Movies Reviews	
Large Movie Review Dataset	http://ai.stanford.edu/~amaas/data/sentiment/
Web data: Amazon movie reviews	http://snap.stanford.edu/data/web-Movies.html
Sentiment Analysis on Movie Reviews	https://www.kaggle.com/c/sentiment-analysis-on-movie-reviews/data.
Movie Review Data	http://www.cs.cornell.edu/People/pabo/movie-review-data/
Twitter	
Sentiment140	http://help.sentiment140.com/for-students
Twitter Sentiment Analysis Dataset	http://thinknook.com/twitter-sentiment-analysis-training-corpus-dataset-2012-09-22/
Twitter Data set for Arabic Sentiment Analysis Data Set	https://archive.ics.uci.edu/ml/datasets/Twitter+Data+set+for+Arabic+Sentiment+Analysis
TASS 2013 Corpus	http://www.daedalus.es/TASS2013/corpus.php