Inter-Annotator Agreement in Sentiment Analysis: Machine Learning Perspective

Victoria Bobicev
Department of Informatics and Systems Engineering
Technical University of Moldova
victoria.bobicev@ia.utm.md

Marina Sokolova
IBDA@Dalhousie University, Halifax, NS, Canada, and
University of Ottawa, Ottawa, ON, Canada
sokolova@uottawa.ca

Abstract

Manual text annotation is an essential part of Big Text analytics. Although annotators work with limited parts of data sets, their results are extrapolated by automated text classification and affect the final classification results. Reliability of annotations and adequacy of assigned labels are especially important in the case of sentiment annotations. In the current study we examine inter-annotator agreement in multi-class, multi-label sentiment annotation of messages. We used several annotation agreement measures, as well as statistical analysis and Machine Learning to assess the resulting annotations.

1 Introduction

Automated text analytics methods rely on manually annotated data while building their heuristics. Although manually annotated texts comprise a small part of a data set, learning algorithms extrapolate the results to the remaining part of the data set and beyond it. At the same time, development of annotation schemes and methods and their implementation attract considerably less research attention than building electronic resources or development of Machine Learning techniques. In this work, we study multi-class sentiment annotation of a *new* data set. The texts have been collected from an online health forum; three annotators annotated each text. Annotators could assign each text with one or more sentiment labels; 4 labels were predefined (*facts, gratitude, encouragement, confusion*) and could be supplemented by other sentiment labels, if annotators deemed the four given labels insufficient to cover sentiments found in the text. Our results show that metrics estimating inter-annotator agreement can be effectively used in deciding of sentiment categories and establishing annotation protocols. We apply Machine Learning (ML) techniques to compare human and automated recognition of sentiment labels.

2 Related Work

Inter-annotator agreement of multi-class sentiment annotations was analyzed in (Steinert, 2017). The author worked with 3255 German documents collected from social networks. Texts were 50 words in average, thus limiting topics and sentiments conveyed by individual messages. Six annotators annotated the documents with one of the following labels: Neg, Neut, Pos, No Sent, Undecided, and Irrelevant. Each document was annotated by three participants. Cohen kappa was used to estimate inter-annotator agreements between each pair of all the six annotators. The best kappa was 0.747 and the worst kappa was 0.480. The final message labels were selected by a majority voting algorithm.

Manual annotations and automated annotations were compared by Emi Ishita et al. (2010). The authors worked with 1,783 sentences collected from written statements of hearings held by a U.S. Senate Committee. Four annotators could annotate a sentence into 10 possible labels; each sentence obtained 3 labels on average; the obtained agreement was 0.30. Then a k-NN algorithm was applied to classify the sentences; it obtained F-score = 0.4. Then human annotations were evaluated in the same way as the automated classification, i.e. obtaining F measures against the final annotation called “ground truth”. The best F-measure for human annotation was 0.70.

Proceedings of Recent Advances in Natural Language Processing, pages 97–102,
Varna, Bulgaria, Sep 4–6 2017.
https://doi.org/10.26615/978-954-452-049-6_015

600 sentences collected from English-language Spine-health forum were manually annotated by 60 Master's students, who were not health professionals (Melzi et al, 2014). Annotators used 6 basic emotions (Ekman, 1992): anger, disgust, fear, joy, sadness and surprise. Each sentence was annotated by two annotators. The inter-annotation coefficient Kappa was 0.26. 150 sentences from the same corpus were annotated by two health professionals. The agreement between health professional annotators and non-professionals was moderate, 0.46. The authors nevertheless proceeded with machine learning experiments; their best F-score was equal to 0.65.

150 topics from 115 documents (from the Blog Track at TREC 2008) were annotated in (Bermingham and Alan, 2009). An average of 3.6 annotators worked with each topic. Annotators' agreement was evaluated by Krippendorf alpha (Hayes and Krippendorf, 2007). Unlike Cohen and Fleiss kappa, the measure can assess agreement among a variable number of annotators and accepts non-annotated examples. The value of 0.4219 was obtained for sentence-level annotation for 5-class annotation; the authors considered such agreement as moderate. Bermingham and Alan emphasized necessity of further studies of different levels of annotations, i.e., sentence, paragraph and document levels.

3 The Data Set and Annotation

The data set. To build this data set, we collected messages posted on Introduction and IVF/FET/IUI Cycle Buddies sub-forums of InVitroFertilization.ca forum[1]. The structure of the posted discussions is similar to those posted on other online forums: a participant starts discussion by posting the first message; other participants join discussion by replying on the initial post or the following messages, thus creating coherent online conversations. The Introduction sub-forum contained 2,913 discussions, whereas the IVF/FET/IUI Cycle Buddies contained 3,771 discussions. The number of messages in a single discussion varied considerably: from hundreds to only a few messages.

Due to budgetary limitations, we restricted the number of annotated discussions. We selected 65 medium length discussions, with 10-20 posts in each discussion. Those discussions yielded in total 1000 posts. The messages were comparatively long, 126 words on average.

Annotation procedure. We adopted four labels from a label set proposed in (Sokolova and Bobicev, 2013); the label set was further analyzed in (Navindgi et al., 2016). We used three sentiment labels: ***confusion***, ***encouragement*** and ***gratitude*** and the label ***facts*** for neutral factual information. We discarded the label ***endorsement*** that was created as combination of two labels ***facts*** and ***encouragement*** (Bobicev et al., 2015). Navindgi et al. (2016) demonstrated that this label confused machine learning algorithms and was not recognized properly.

As one message often conveyed > 1 sentiment, the resulting sentiment mosaic posed challenges to both one label and multi-labeled post annotation:

- one label annotation can be directly mapped into a multi-class classification problem; however, deciding on one sentiment label will artificially restrict annotators' choices, thus, leaving side important information about message sentiment and their annotation.
- multi-labeled annotation represents diversity of sentiments in individual posts; however, mapping multitudes of labels into multi-label machine learning is not a trivial task.

In this study, we have decided to combine the two approaches: first, allow multiple labels for one post; then generalize the received annotations into one label per post.

Preliminary text annotation. In the annotation guidelines we presented three sentiment labels and one for neural/factual information as default labels and then suggested that for each message annotators can indicate as many sentiments as they considered appropriate. Although the annotators were allowed to attach any number of any labels, in many cases they assigned only one predefined label per post: approx. 85% of the assigned labels were from the predefined set. New labels suggested by the annotators are reported in Table 1.

The final text annotation. To choose one label for each post based on all the labels assigned by annotators, we had to resolve the following situations: (1) All three annotators assigned the same label for a post; no other labels were assigned; that ideal case happened in 326 posts. (2) All three annotators assigned the same label for a post; however, other, non-matching labels were assigned too; in this case, the label indicated by all three annotators was selected as the final one: 297

[1] www.ivf.ca

posts. (3) All three annotators assigned same two labels for a post; no other labels were assigned; it happened for 26 posts; one annotator indicated the importance of the labels; we used this information and assigned the label she indicated first. (4) Of all labels attached to the post by all annotators only one label was used twice; this label was selected as the final one: 214 posts (5) Two or even three labels were used twice in annotation of the same post; in this case we kept the label marked as the most important; it happened for 95 posts; in four cases the most important label was not repeated by other annotators; we did not use these posts in the experiments. (6) All the annotators attached different labels to a post. This happened for 14 posts. Those posts remained ambiguous without the final label.

Table 1: New sentiments of 1st, 2nd and 3rd annotators.

Sentiment	1st	2nd	3rd
support		473	
cheering		158	
worry		124	
uncertainty		115	
compassion	53	96	17
hope	37	238	19
optimism		125	
dislike		92	
excitement	20		
concern		81	
sadness		38	
joy			5
happiness		36	3
disappointment	13		2
sadness	9		12
frustration	7		

Figure 1 reports labels assigned by annotators after we merged the additional labels. The first annotator tended to use more neutral labels whilst the second annotator attached more labels with sentiments. This problem was discussed in (Melzi et al., 2014) as health forums are about health problems, diagnosis, treatment, etc. Thus, some annotators, by empathy, associated negative emotion to factual information about diseases, symptoms and diagnoses.

In general, the first annotator attached fewer labels (average labels per post is 1.17; 83% of posts annotated with only one label); the second and the third annotators attached more labels (average labels per post is 1.35, 68% of posts annotated with only one label).

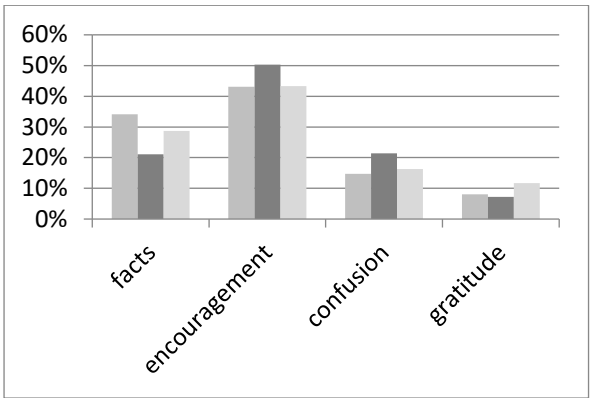

Figure 1: The ratio of labels used by all three annotators in the order: 1st, 2nd, 3rd.

4 Assessment of Inter-Annotator Agreement

Several inter-annotators metrics were proposed for inter-annotator agreement evaluation.

Per cent of agreement is the easiest and the most straightforward measure. The measure is often criticized (it does not differentiate between categories) but still it gives a basic approximation of annotators' agreement. As we have multiple labels per post, inter-annotator agreement was calculated for each label separately and the average for these four labels. The same situation is for three annotators. The pairwise agreement calculated for each pair of the annotators and the average are presented in Table 2.

Table 2: Per cent inter-annotator agreement for the four labels and three annotators. A1, A2, A3 are the annotators; the labels are: lab1 (encouragement), lab2 (facts), lab3 (confusion), lab4 (gratitude).

	lab1	lab2	lab3	lab4
A1vs A2	73%	72%	85%	88%
A2vs A3	73%	70%	83%	93%
A1vs A3	78%	74%	81%	87%
average	74.6%	71.8%	82.7%	89.1%

The average agreement is 79.6%; it shows a strong agreement. Agreement between pairs of the three annotators is almost uniform; we cannot say that some are better than others.

Comparing agreement per labels we see that lesser agreement was reached on ***encouragement*** and ***facts***. This may be explained by the specific of the text: participants often asked to tell success stories to encourage them fighting with health issues. Thus, factual stories of illness and following

recovery were frequently perceived as encouragement.

Inter-annotator agreement is often evaluated by Cohen's kappa (κ), Fleiss' kappa (K), Krippendorf's alpha (α) (Artstein, Poesio, 2008). Cohen's kappa and Fleiss kappa are Chance-Corrected Coefficients which means that they measure above chance agreement. The general formula for these coefficients is:

$$k = (Ao - Ae)/(1 - Ae) \qquad (1)$$

where *Ao* is observed agreement and *Ae* is expected agreement by chance if the annotators pick the labels randomly.

Cohen *k*. In case of Cohen *k* the expected agreement *Ae* is calculated basing on the assumption that random assignment of categories to the items is governed by prior distributions that are unique to each coder, and which are observed from their actual distribution. Cohen *k* is applied to two annotators only, thus we calculated it for each pair of our annotators (Table 3)

Table 3: Cohen's ***k*** for the annotation

	lab1	lab2	lab3	lab4
A1vs A2	0.48	0.41	0.50	0.47
A1vs A2	0.46	0.34	0.51	0.56
A1vs A2	0.53	0.40	0.49	0.44
average	0.49	0.38	0.50	0.49

The average Cohen's k = 0.46 shows moderate agreement.

Fleiss *k* is generalization to more than two annotators; expected agreement is calculated on the basis on the assumption that random assignment of categories to items, by any annotator, is governed by the distribution of items among categories in the actual world (Table 4).

Table 4: Fleiss ***k*** for the annotation.

	lab1	lab2	lab3	lab4	average
Fleiss ***k***	0.48	0.38	0.49	0.48	0.46
observed	0.75	0.72	0.83	0.89	0.80
expected	0.51	0.55	0.66	0.79	0.63

Krippendorf's alpha (α) is an agreement coefficient based on assumptions that expected agreement is calculated by looking at the overall distribution of judgments without regard to which coders produced these judgments. It applies to multiple coders, and it allows missing values. As in previous cases α is calculated for each label and then averaged in Table 5.

Table 5: Krippendorf's alpha (α)

	lab1	lab2	lab3	lab4	average
α	0.48	0.38	0.49	0.48	0.46

The average agreement is equal to 0.46 in all calculated metrics which is similar to 0.48 reported in (Bobicev, Sokolova, 2017) and is considered as moderate in (Artstein, Poesio, 2008).

In the current study, there was no considerable difference between the results of the inter-annotator metrics. Nevertheless, we conjecture that pairwise metrics can serve for the detection of the best annotator. For example, Nowak and Roger (2010) identified the best and the worst annotators (among 11 annotators) by calculating correlation between their labels and the final labeled set.

On the other hand, we can calculate annotator agreement per label and then use the result as an indicator of the label identification difficulty: labels with the best agreement are the easiest to detect and, vice versa, the worst agreement shows labels with the most problems for the annotators. Finally, annotator agreement can be used in analysis of the label sets and annotation instructions in order to improve annotation process and resolve annotation difficulties.

5 Experiments

We worked with the following sets of text:

(1) 970 posts, each post identified with one label;

(2) 326 posts with exact match of the labels assigned by the three annotators;

(3) 297 posts with a principal match: all 3 annotators agreed on 1 label but other labels were added;

(4) 214 posts with a partial agreement: two annotators selected one label which was considered the final one;

The complete set of 970 messages has the following distribution of labels: ***confusion***: 146, ***encouragement***: 494, ***gratitude***: 69, ***facts***: 261.

To estimate reliability of annotations (see Sec 4), we have run ML experiments on sets (1) – (3). We used several sets of features:

1. BOW – Bag of Words. All words from the whole corpus with the frequency at least 2; 3497 features.

2. SS – SentiStrength (Thelwall et al., 2012). All terms from SentiStrength lexicon which appear in our corpus.
3. SWN – SentiWordNet (Esuli, Sebastiani, 2006). All terms from SentiWordNet lexicon which appear in our corpus.
4. DM – DepecheMood (Staiano, Guerini, 2014). All terms from DepecheMood lexicon which appear in our corpus.
5. HA – HealthAffect (Sokolova, Bobicev, 2013). All terms from HealthAffect lexicon which appear in our corpus.

Table 6: Results of the 1st set of the experiments.

experiment	**F**	**Features**	**algorithm**
1	0.794	BOW	NB Multinomial
2	0.504	BOW	SVM
3	0.445	SWN	NB Multinomial
4	0.571	BOW	NB Multinomial

Note that SS, SWN, DM, HA showed reliable results in sentiment analysis studies (Bobicev et al., 2015). We applied Naive Bayes (NB), NB Text, NB multinomial, SVM machine learning algorithms in our experiments. To select the best results, we used 10-fold cross-validation and computed F-score (F). We used the majority class baseline. We run four experiments using the created sub-corpora (Table 6):

(1) 326 posts with exact match; baseline=0.536.
(2) 297 posts with agreement; baseline=0.362.
(3) 214 posts with partial agreement; baseline=0.283.
(4) All the 970 posts with one label; baseline=0.344.

The best result F=0.794 was obtained for the sub-corpus with exact match. For the whole corpus the best F-measure=0.571 was too obtained on the BOW feature set.

Table 7: Results of the 2nd set of the experiments on the selected set of features.

experiment	**F**	**algorithm**
1	0.797	NB Multinomial
2	0.628	NB Multinomial
3	0.527	NB Multinomial
4	0.646	NB Multinomial

Table 8: Classification per label

Class	F-Measure
confusion	0.562
facts	0.535
gratitude	0.441
encouragement	0.759

Feature selection Feature sets consisting of all words from the texts or from lexicons were comparatively large. We used Correlation-based Feature Subset Selection (Hall, 1999) which evaluates the worth of a subset of features by considering the individual predictive ability of each feature along with the degree of redundancy between them. We created a set of selected features from all 4 lexicons' merged features and have run the same 6 experiments on them. The results are presented in Table 7. Feature selection improved the learning results: F measure for the whole set is 0.652 which is much better than 0.571 on BOW features.

F-measures for the individual labels are presented in Table 8. We used the feature selection results obtained on all the 970 posts with 1 label, i.e. experiment # 4 in Table 7. ***Encouragement*** was classified significantly better than other categories: F-score = 0.759; the lowest F-score= 0.441 was obtained for classification of ***gratitude***.

Those results may be explained by imbalanced distribution of categories: almost half of posts were labeled by ***encouragement*** (i.e., 494 posts out of 970 posts) while ***gratitude*** had been assigned to the smallest number of posts (i.e., 69 out of 970). Thus, the automate method had enough training data to recognize ***encouragement***, but not enough training data for recognition of ***gratitude***.

Table 9: Inter-annotator agreement calculated for automated classification and final labels.

label	Per cent Agreement	Cohen kappa
confusion	85.9	0.478
facts	76.7	0.380
gratitude	93.2	0.405
encouragement	74.5	0.489

To compare automated classification and human annotation we calculated agreement between the results of automated analysis and the final set of labels used in the experiments. Cohen kappa is the worst for ***facts*** (0.380) and the best for ***encouragement*** (0.489) (Table 9). Thus, humans and

automated methods disagree more when they assign ***facts*** than when they assign ***encouragement.***

6 Conclusions and Future Work

We have presented a study of multi-class sentiment annotation. We worked with a *new* data set of 970 texts collected from a health-related forum. To estimate the quality of annotations, we have applied several inter-annotation agreement metrics. We have shown how those metrics can be used in evaluation of sentiment categories and annotation schemes. Finally, we applied Machine Learning techniques to compare automated classification and human annotation of the same data.

Our future work will expand current studies to new data sets. We aim to investigate various protocols and procedures of generalization of sentiment annotations, and how those procedures affect Machine Learning sentiment classification.

References

Ron Artstein, Massimo Poesio. 2008 Inter-coder agreement for computational linguistics. Computational Linguistics Journal Volume 34 Issue 4, pages 555-596. doi: 10.1162/coli.07-034-R2

Adam Bermingham, Alan F. Smeaton. 2009 A study of inter-annotator agreement for opinion retrieval. In: SIGIR 2009 - The 32nd Annual ACM SIGIR Conference, pages 784-785.

Victoria Bobicev, Marina Sokolova and Michael Oakes. 2015. What Goes Around Comes Around: Learning Sentiments in Online Medical Forums. *Cognitive Computation*, 7(5): 609-621. http://dx.doi.org/10.1007/s12559-015-9327-y.

Victoria Bobicev, Marina Sokolova. 2017. Confused and Thankful: Multi-label Sentiment Classification of Health Forums. In: Mouhoub M., Langlais P. (eds) Advances in Artificial Intelligence. AI 2017. DOI: 10.1007/978-3-319-57351-9_33

Paul Ekman. 1992. An argument for basic emotions. Cognition and Emotion, vol. 6(3-4), pp. 169–200.

Andrea Esuli and Fabrizio Sebastiani. 2006. SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining, In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06), pages 417-422.

Mark A. Hall. 1999. Correlation-based Feature Selection for Machine Learning, Ph.D. thesis, University of Waikato.

Andrew F. Hayes and Klaus Krippendorff. 2007. Answering the Call for a Standard Reliability Measure for Coding Data. Communication Methods and Measures Journal Volume 1, Issue 1, pp. 77-89.

Emi Ishita, Douglas W. Oard, Kenneth R. Fleischmann, An-Shou Cheng, Thomas Clay Templeton. 2010. Investigating multi-label classification for human values. Proceedings of the Association for Information Science and Technology. http://dx.doi.org/10.1002/meet.14504701116

Soumia Melzi, Amine Abdaoui, Jérôme Azé, Sandra Bringay, Pascal Poncelet, Florence Galtier. 2014. Patient's rationale: Patient Knowledge retrieval from health forums. eTELEMED 2014 : The Sixth International Conference on eHealth, Telemedicine, and Social Medicine.

Soumia Melzi, Amine Abdaoui, Jérôme Azé, Sandra Bringay, Pascal Poncelet, Florence Galtier. 2014. Patient's rationale: Patient Knowledge retrieval from health forums. eTELEMED 2014 : The Sixth International Conference on eHealth, Telemedicine, and Social Medicine.

Stefanie Nowak, Stefan Roger. 2010. How reliable are annotations via crowdsourcing? a study about inter-annotator agreement for multi-label image annotation. In: Proceedings of the international conference on Multimedia information retrieval - MIR'10, p. 557.

Marina Sokolova and Victoria Bobicev. 2013. What Sentiments Can Be Found in Medical Forums? In Galia Angelova, Kalina Bontcheva, Ruslan Mitkov: Recent Advances in Natural Language Processing, RANLP 2013, Bulgaria.

Jacopo Staiano, Marco Guerini. 2014. DepecheMood: a Lexicon for Emotion Analysis from Crowd-Annotated News. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pages 427-433. http://www.anthology.aclweb.org/P/P14/

Kai Steinert. 2017 Collaborative Web-Based Short Text Annotation with Online Label Suggestion. MA Thesis, TU Darmstadt.

Mike Thelwall, Kevan Buckley, Georgios Paltoglou. 2012. Sentiment strength detection for the social Web, Journal of the American Society for Information Science and Technology, 63(1), 163-173. doi>10.1002/asi.21662.