A Practical Guide to Sentiment Annotation: Challenges and Solutions

Saif M. Mohammad
National Research Council Canada
saif.mohammad@nrc-cnrc.gc.ca

Abstract

Sentences and tweets are often annotated for sentiment simply by asking respondents to label them as positive, negative, or neutral. This works well for simple expressions of sentiment; however, for many other types of sentences, respondents are unsure of how to annotate, and produce inconsistent labels. In this paper, we outline several types of sentences that are particularly challenging for manual sentiment annotation. Next we propose two annotation schemes that address these challenges, and list benefits and limitations for both.

1 Introduction

Clear and simple instructions are crucial for obtaining high-quality annotations. This is true even for seemingly simple annotation tasks, such as sentiment annotation, where one is to label instances as positive, negative, or neutral. For word annotations, researchers have often framed the task as 'is this word positive, negative, or neutral?' (Hu and Liu, 2004), 'does this word have associations with positive, negative, or neutral sentiment?' (Mohammad and Turney, 2013), or 'which word is more positive?'/'which word has a greater association with positive sentiment' (Kiritchenko et al., 2016; Kiritchenko and Mohammad, 2016b). Similar instructions are also widely used for sentence-level sentiment annotations—'is this sentence positive, negative, or neutral?' (Rosenthal et al., 2015; Rosenthal et al., 2014; Mohammad et al., 2016a; Mohammad et al., 2015). We will refer to such annotation schemes as *the simple sentiment questionnaires*. On the one hand, this characterization of the task is simple, terse, and reliant on the intuitions of native speakers of a language (rather than biasing the annotators by providing definitions of what it means to be positive, negative, and neutral). On the other hand, the lack of specification leaves the annotator in doubt over how to label certain kinds of instances—for example, sentences where one side wins against another, sarcastic sentences, or retweets.

A different approach to sentiment annotation is to ask respondents to identify the target of opinion, and the sentiment towards this target of opinion (Pontiki et al., 2014; Mohammad et al., 2015; Deng and Wiebe, 2014). We will refer to such annotation schemes as *the semantic-role based sentiment questionnaires*. This approach of sentiment annotation is more specific, and more involved, than the simple sentiment questionnaire approach; however, it too is insufficient for handling several scenarios. Most notably, the emotional state of the speaker is not under the purview of this scheme. Many applications require that statements expressing positive or negative emotional state of the speaker should be marked as 'positive' or 'negative', respectively. Similarly, many applications require statements that describe positive or negative events or situations to be marked as 'positive' or 'negative', respectively. Instructions for annotating opinion towards targets do not specify how such instances are to be annotated, and worse still, possibly imply that such instances are to be labeled as neutral.

In this paper, we present a list of sentence types that are especially challenging for sentiment annotation. Next, we propose two annotation schemes that

Proceedings of NAACL-HLT 2016, pages 174–179,
San Diego, California, June 12-17, 2016.

address these challenges: (1) a simple sentiment annotation questionnaire with more precise annotation directions and some additional label categories; and (2) a semantic-role based questionnaire with additional questions to account for the speaker's emotional state and descriptions of valenced events.

Aspects of annotation that are not specific to sentiment, such as good practices in crowdsourcing, how to aggregate information from multiple annotators, and how to automatically detect and discard poor annotations are beyond the scope of this paper; we refer the readers to Lease (2011), Hsueh et al. (2009), and Mohammad and Turney (2013) for that. Methods for obtaining real-valued sentiment scores are also not covered in this paper; we refer the reader to Kiritchenko and Mohammad (2016a) and Kiritchenko et al. (2014) for the use of best–worst scaling to obtain reliable real-valued sentiment associations. See Mohammad (2016) for a survey on sentiment and emotion datasets.

2 Types of Instances that are Difficult to Annotate for Sentiment

There exist several types of sentences that are particularly challenging to annotate for sentiment. Some of the more notable ones are listed below:

- *Speaker's emotional state:* The speaker's emotional state may or may not have the same polarity as the opinion expressed by the speaker. For example, a politician's tweet can imply both a negative opinion about a rival's past indiscretion, and a joyous mental state as the news will impact the rival adversely.

- *Success or failure of one side w.r.t. another*: Often sentences describe the success or failure of one side w.r.t. another side—for example, *'Yay! France beat Germany 3–1', 'Supreme court judges in favor of gay marriage'*, and *'the coalition captured the rebels'*. If one supports France, gay marriage, and the coalition, then these events are positive, but if one supports Germany, marriage as a union only between man and woman, and the rebels, then these events can be seen as negative.
  Also note that the framing of an event as the success of one party (or as the failure of another party) does not automatically imply that the speaker is expressing positive (or negative) opinion towards the mentioned party. For example, when Finland beat Russia in ice hockey in the 2014 Sochi Winter Olympics, the event was tweeted around the world predominantly as "Russia lost to Finland" as opposed to "Finland beat Russia". This is not because the speakers were expressing negative opinion towards the Russian team, but rather simply because Russia, being the host nation, was the focus of attention and traditionally Russian hockey teams have been strong.

- *Neutral reporting of valenced information*: If the speaker does not give any indication of her own emotional state but describes valenced events or situations, then it is unclear whether to consider these statements as neutral unemotional reporting of developments or whether to assume that the speaker is in a negative emotional state (sad, angry, etc.). Example:

  The war has created millions of refugees.

- *Sarcasm and ridicule*: Sarcasm and ridicule are tricky from the perspective of assigning a single label of sentiment because they can often indicate positive emotional state of the speaker (pleasure from mocking someone or something) even though they have a negative attitude towards someone or something.

- *Different sentiment towards different targets of opinion*: The speaker may express opinion about multiple targets, and sentiment towards the different targets might be different. The targets may be different people or objects (for example, an iPhone vs. an android phone), or they may be different aspects of the same entity (for example, quality of service vs. quality of food at a restaurant).

- *Precisely determining the target of opinion*: Sometimes it is difficult to precisely identify the target of opinion. For example, consider:

  Glad to see Hillary's lies being exposed.

  It is unclear whether the target of opinion is 'Hillary', 'Hillary's lies', or 'Hillary's lies being exposed'. One reasonable interpretation is that

positive sentiment is expressed about 'Hillary's lies being exposed'. However, one can also infer that the speaker has a negative attitude towards 'Hillary's lies' and probably 'Hillary' in general. It is unclear whether annotators should be asked to provide all three opinion–target pairs or only one (in which case, which one?).

- *Supplications and requests*: Many tweets convey positive supplications to God or positive requests to people in the context of a (usually) negative situation. Examples include:

  May god help those displaced by war. Let us all come together and say no to fear mongering and divisive politics.

- *Rhetorical questions*: Rhetorical questions can be treated simply as queries (and thus neutral) or as utterances that give away the emotional state of the speaker. For example, consider:

  Why do we have to quibble every time?

  On the one hand, this tweet can be treated as a neutral question, but on the other hand, it can be seen as negative because the utterance betrays a sense of frustration on the part of the speaker.

- *Quoting somebody else or re-tweeting*: Quotes and retweets are difficult to annotate for sentiment because it is often unclear and not explicitly evident whether the one who quotes (or retweets) holds the same opinions as that expressed by the quotee.

The challenges listed above can be addressed to varying degrees by providing instructions to the annotators on how such instances are to be labeled. However, detailed and complicated instructions can be counter-productive as the annotators may not understand or may not have the inclination to understand the subtleties involved.

3 Proposed Annotation Schemes

Two annotation schemes that address many of the challenges laid out above are presented below. The benefits and limitations of both are outlined. The goal here is not to suggest that these are the only ways to annotate for sentiment, but rather to encourage further thought and improved proposals for sentiment annotation (that may or may not be inspired by the questionnaires shown below). Note also that the precise formulation of the sentiment questionnaire should be guided by the specific needs of the application at hand.

3.1 A Simple Sentiment Questionnaire

A simple sentiment questionnaire that addresses many of the challenges listed in Section 2 is presented below:

PROPOSED SIMPLE SENTIMENT QUESTIONNAIRE

What kind of language is the speaker using?

1. the speaker is using positive language, for example, expressions of support, admiration, positive attitude, forgiveness, fostering, success, positive emotional state
2. the speaker is using negative language, for example, expressions of criticism, judgment, negative attitude, questioning validity/competence, failure, negative emotion
3. the speaker is using expressions of sarcasm, ridicule, or mockery
4. the speaker is using positive language in part and negative language in part
5. the speaker is neither using positive language nor using negative language

Notes:

- A good response to this question is one that most people will agree with. For example, even if you think that sometimes the language can be considered negative, if you think most people will consider the language to be positive, then select the positive language option.
- Agreeing or disagreeing with the speaker's views should not have a bearing on your response. You are to assess the language being used (not the views). For example, given the tweet, 'Evolution makes no sense', the correct answer is 'the speaker is using negative language' since the speaker's words are criticizing or judging negatively something (in this case the theory of evolution). Note that the answer is not contingent on whether you believe in evolution or not.

Benefits and Limitations. This questionnaire groups the speaker's emotional state, speaker's opinion, and description of valenced events all into one category and aims simply to determine the dominant sentiment inferable from the sentence. The phrases 'positive language' and 'negative language' encourage respondents to focus on the language itself as opposed to assigning sentiment based on event outcomes that are beneficial to them. For example, 'Yay! France beat Germany 3–1' will be marked as positive because the speaker is using the positive expression 'Yay!'. The 'Russia lost to Finland' example (described earlier in Section 2), may be difficult to annotate with respect to the opinion of the speaker towards the Russian team, but the framing of the event as a loss is easily identified as negative language. Other instances where one side benefits over another, but the text itself does not use positive or negative language can be labeled as option 5. Sarcasm, ridicule, and mockery are included as a separate option (in addition to option 2) so that respondents do not have to struggle with the decision of whether to mark such instances as positive or negative. Downstream applications can make use of all of these annotations as is appropriate for them: for example, by treating tweets labeled sarcasm and ridicule differently from the other positive and negative sentences, or by marking them as negative.

Instances with different sentiment towards different targets of opinion can be marked with option 4. Supplications and requests that convey a sense of fostering and support can be marked as positive. On the other hand, rhetorical questions that betray a sense of frustration and disappointment can be marked as negative. Thus this simple questionnaire addresses many of the issues raised in the earlier section. However, a limitation of this approach is that it does not produce as nuanced a set of annotations as those that can be obtained from the questionnaire shown ahead. Note, however, that the simplcity of the questionnaire entails low annotation costs and no special educational requirements. Mohammad et al. (2016b) used this questionnaire to annotate a set of tweets for sentiment via crowdsourcing. The dataset, which is also annotated for stance, is made freely available.[1]

[1] www.saifmohammad.com/WebPages/StanceDataset.htm

3.2 A Semantic-Role Based Sentiment Questionnaire

A sentiment questionnaire that includes questions about the target of opinion, as well as additional questions such as those that address the speaker's emotional state, is presented below:

PROPOSED SEMANTIC-ROLE BASED SENTIMENT QUESTIONNAIRE

Q1. From reading the text, the speaker's emotional state can best be described as:

- *positive state*: there is an explicit or implicit clue in the text suggesting that the speaker is in a positive state, i.e., happy, admiring, relaxed, forgiving, etc.
- *negative state*: there is an explicit or implicit clue in the text suggesting that the speaker is in a negative state, i.e., sad, angry, anxious, violent, etc.
- *both positive and negative, or mixed, feelings*: there is an explicit or implicit clue in the text suggesting that the speaker is experiencing both positive and negative feelings
- *unknown state*: there is no explicit or implicit indicator of the speaker's emotional state

Q2. From reading the text, identify the entity towards which opinion is being expressed or the entity towards which the speaker's attitude can be determined.

This entity is usually a person, object, company, group of people, or some such entity. We will call this the PRIMARY TARGET OF OPINION (PTO). For example, if the text criticizes certain actions or beliefs of a person (or group of persons), then that person or group is the PTO. If the text mocks people who do not believe in evolution, then the PTO is 'people who do not believe in evolution'. If the text questions or mocks evolution, then the PTO is 'evolution'. If you cannot determine sentiment/attitude of the speaker towards a person, group, or object, but you can identify sentiment/attitude towards an action or event, then consider that action or event as the PTO. If there are more than one targets of opinion, then select that target towards which sentiment is stronger.

NOTE: Where possible, copy and paste the primary target of opinion from the text. If the target can be referred to in different ways, for example, Barack Obama, Obama, Obamaaa, #obama, @obama, President, he, etc., copy and paste the snippet from the text showing how the speaker has referred to the target.

Q3. What best describes the speaker's attitude, evaluation, or judgment towards the primary target of opinion (PTO)? If the whole text is a quote from somebody else (original author) and there is no indication of speaker's attitude, then answer below considering the original author as the speaker.

- *positive*: there is an explicit or implicit clue in the text suggesting that the speaker's attitude or judgment of the PTO is positive (speaker is appreciative, thankful, excited, optimistic, or inspired by the primary entity)
- *negative*: there is an explicit or implicit clue in the text suggesting that the speaker's attitude or judgment of the PTO is negative (speaker is critical, angry, disappointed in, pessimistic, expressing sarcasm about, or mocking the primary entity)
- *mixed*: there is an explicit or implicit clue in the text suggesting that the speaker's attitude or judgment of the PTO is both positive and negative
- *unknown*: there is no explicit or implicit clue indicating that the speaker feels positively or negatively

Q4. What best describes the sentimental impact of the primary target of opinion (PTO) on most people?

- *positive*: the PTO is considered predominantly positive
- *negative*: the PTO is considered predominantly negative
- *mixed (both positive and negative)*: some aspects of the PTO are positive and some are negative
- *mixed (opposing sides)*: the PTO is considered positive by a large group of people AND is considered negative by another large group of people
- *no sentiment*: there is no clear sentiment associated with the PTO

Examples:

- For Q1:
  - Text: *Mugabe killed millions during his rule*
    Answer: unknown state (since there is no clue about the emotional state of the speaker)
  - Text: *Arggh! When will politicians learn to govern?*
    Answer: negative state (since there is sufficient indication that the speaker is frustrated)
- For Q2:
  - Text: *Sorry to see Mugabe kill so many civilians.*
    Answer: Mugabe
  - Text: *When will they stop killing babies in the womb?*'
    Answer: 'they'
- For Q3:
  - Text: *Sorry to see Mugabe kill so many civilians.*
    Answer: negative (We can infer that the speaker has negative sentiment toward Mugabe.)
  - Text: *We need a diplomat like Kissinger*
    Answer: positive (We can infer that the speaker has a positive attitude towards Kissinger.)
- For Q4:
  - Text: *Hillary has to answer for Benghazi.*
    Answer: mixed (opposing sides) (The speaker is expressing negative sentiment towards Hillary, but there are many who view Hillary favorably.)
  - Text: *The war has displaced millions*
    Answer: negative (this event is predominantly negative)

Benefits and Limitations: This detailed questionnaire with questions for the speaker's emotional state, the target of opinion, opinion towards the target, and general opinion (not the speaker's opinion) towards the target provides a rich cross-section of information that can be used by many downstream applications. However, a limitation of this questionnaire is its complexity—annotators (especially on crowdsourcing platforms) might find it difficult to distinguish the subtle difference between Q1, Q3, and Q4. Additionally, even though Q2 is framed with an eye on challenges regarding the identification of the target of opinion, it is difficult to completely address the associated issues. The target may not be explicitly mentioned in the text, it may be mentioned multiple times and in different ways, or it may be referred to via hypernyms, hyponyms, meronyms, and holonyms. It is advisable to first train the annotators on a small set of instances before proceeding to annotate large amounts of data.

4 Summary

We outlined several types of sentences that are particularly challenging for manual sentiment annotation. They include sentences describing success (or failure) of one side over another, sentences expressing sarcasm or ridicule, sentences expressing differing sentiment towards multiple entities, supplications, requests, and rhetorical questions. We then presented two different questionnaires that provide clear instructions on how such instances are to be annotated. The first questionnaire, is simple and terse, and thus it is easy and inexpensive to answer. The second questionnaire is markedly more involved, and thus requires more training; however, it provides a plethora of sentiment-related information that can be used in many downstream applications. Apsects of the proposed questionnaires may not be appropriate for all applications. Practitioners are encouraged to tweak the questionnaires as per the needs of the application at hand.

Acknowledgments

Many thanks to Svetlana Kiritchenko for helpful discussions.

References

Lingjia Deng and Janyce Wiebe. 2014. Sentiment propagation via implicature constraints. In *EACL*, pages 377–385.

Pei-Yun Hsueh, Prem Melville, and Vikas Sindhwani. 2009. Data quality from crowdsourcing: a study of annotation selection criteria. In *Proceedings of the NAACL HLT 2009 workshop on active learning for natural language processing*, pages 27–35. Association for Computational Linguistics.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pages 168–177, New York, NY, USA. ACM.

Svetlana Kiritchenko and Saif M. Mohammad. 2016a. Capturing reliable fine-grained sentiment associations by crowdsourcing and best–worst scaling. In *Proceedings of The 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, San Diego, California.

Svetlana Kiritchenko and Saif M. Mohammad. 2016b. Sentiment composition of words with opposing polarities. In *Proceedings of The 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, San Diego, California.

Svetlana Kiritchenko, Xiaodan Zhu, and Saif M. Mohammad. 2014. Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research*, 50:723–762.

Svetlana Kiritchenko, Saif M. Mohammad, and Mohammad Salameh. 2016. Semeval-2016 task 7: Determining sentiment intensity of english and arabic phrases. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval-2016, San Diego, California.

Matthew Lease. 2011. On quality control and machine learning in crowdsourcing. *Human Computation*, 11:11.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word-emotion association lexicon. 29(3):436–465.

Saif M. Mohammad, Xiaodan Zhu, Svetlana Kiritchenko, and Joel Martin. 2015. Sentiment, emotion, purpose, and style in electoral tweets. *Information Processing and Management*, 51(4):480–499.

Saif M. Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016a. Semeval-2016 Task 6: Detecting stance in tweets. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval '16, San Diego, California.

Saif M. Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko. 2016b. Stance and sentiment in tweets. *Special Section of the ACM Transactions on Internet Technology on Argumentation in Social Media*, Submitted.

Saif M. Mohammad. 2016. Sentiment analysis: Detecting valence, emotions, and other affectual states from text. In Herb Meiselman, editor, *Emotion Measurement*. Elsevier.

Maria Pontiki, Harris Papageorgiou, Dimitrios Galanis, Ion Androutsopoulos, John Pavlopoulos, and Suresh Manandhar. 2014. SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation*, SemEval '14, Dublin, Ireland.

Sara Rosenthal, Alan Ritter, Preslav Nakov, and Veselin Stoyanov. 2014. SemEval-2014 Task 9: Sentiment analysis in Twitter. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 73–80, Dublin, Ireland.

Sara Rosenthal, Preslav Nakov, Svetlana Kiritchenko, Saif Mohammad, Alan Ritter, and Veselin Stoyanov. 2015. SemEval-2015 task 10: Sentiment analysis in Twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation*, SemEval '15, pages 450–462, Denver, Colorado.