Sentiment Classification and Analysis in Twitter

Justin Cosentino
Department of Computer Science
Swarthmore College
Swarthmore, PA 19081
jcosent1@swarthmore.edu

Emanuel Schorsch
Department of Computer Science
Swarthmore College
Swarthmore, PA 19081
eschors1@swarthmore.edu

Abstract

1 Introduction

2 Related Works

3 Methodology

3.1 Data

3.1.1 Twitter Corpus

3.1.2 Subjectivity Lexicon

3.2 Features

3.3 Additive Smoothing

3.4 Sentiment Classifiers

3.4.1 Naive Bayes

Using the Naive Bayes classifier, a given feature vector $\vec{f}$ is assigned the polarity ŝ such that $\hat{s} = \underset{s \in S}{\arg\max} P(s|\vec{f})$. This sentiment classifier is based upon and derived from a fundamental statistical rule called Bayes' law:

$$\hat{s} = \underset{s \in S}{\arg\max} \frac{P(\vec{f}|s)P(s)}{P(\vec{f})}$$

Because $P(\vec{f})$ remains the same value across all possible polarities, it does not assist in determing the value of $\hat{s}$. We then have:

$$\hat{s} = \underset{s \in S}{\arg\max} P(\vec{f}|s)P(s)$$

However, given a feature vector $\vec{f}$ it is very unlikely that we will see this exact feature again. Thus we naively assume that each $\vec{f_i}$ of $\vec{f}$ is independent of all other $\vec{f_i}$. Making this assumption allows us to approximate $P(\vec{f}|s)$:

$$P(\vec{f}|s) \approx \prod_{i=1}^{n} P(f_i|s)$$

Thus, in estimating the probability of the vector $\vec{f}$ by finding the probability of the probabilities pertaining to each individual feature within $\vec{f}$ given the polarity $\hat{s}$, we find the final equation for the Naive Bayes classifier:

$$\hat{s} = \underset{s \in S}{\arg\max} \prod_{i=1}^{n} P(f_i|s)P(s)$$

The maximum likelihood estimate of the probability of each possible sentiment polarity is calculated by taking the count of the number of times a given feature occurs given the polarity over the number of times the count occurs. This is represented by:

$$P(s_i) = \frac{count(s_i, w_j)}{count(w_j)}$$

The probability of each feature is calculated in a similar manner such that

$$P(f_i|s) = \frac{count(f_i, s)}{count(s)}$$

3.4.2 Decision List

A decision list classifier is equivalent to simple case statements or an extended if-else statement. Decision lists generate a set of rules or conditions, for which there is a singly classification associated.

These rules are generated from tagged feature vectors and then scored and ordered based on their associated scores. Similar to the approach used by Yarowsky (CITATION), each pair of feature and value are treated as a rule. In order to generate the best rule for each possible classification, the following equation is used:

$$\left| Log\left(\frac{P(Sense_i|f_i)}{P(Allothersenses|f_i)}\right)\right|$$

Once these rules have been generated from the given feature sets and scored, the decision list creates a list of rules similar to those seen in Figure 1 (MUST CREATE FIGURE WITH EXAMPLE RULES). The decision list will then use these rules as long as their respective scores remain above zero. At this point, the decision list will proceed to classify words based on the most frequent sense of all tweets in the given corpus.

3.4.3 Subjectivity Lexicon

A subjectivity lexicon was used to classify and determine the polarity of tweets. The subjectivity lexicon was acquired from OpinionFinder, a list containing roughly 6,500 words, their polarity, and the strength of their subjectivity (CITATION). This lexicon was then used to determine the polarity of each word within the given tweet or tweet segment. The strength of each word, which was labeled as either strong or weak subjectivity, was ignored for testing.

Individual counts of the number of positive and negative words for each tweet were than kept. If a tweet or tweet segment contained more positive words than negative words, the tweet was then defined as having a positive polarity. If the tweet contained more negative words than positive words, the tweet was determined to have a negative polarity. However, if a tweet contained neither any positive words nor any negative words or if the count of positive and negative words were equal, the tweet was labeled as objective. Within the subjectivity lexicon classifier, words were not labeled as neutral.

Because the subjectivity classifier requires no labeled tweets to be used as training data, the classifier was tested on all tweets within each corpus. However, in order to compare the results of the subjectivity lexicon classifier to the results of our other classifiers, the lexicon was also used to label only the test data used by all other classifiers.

4 Results

5 Analysis

6 Conclusion

6.1 Future Work

7 Text Given

7.1 Electronically-available resources

This description is provided in LaTeX (cs65f12.tex) along with the LaTeX style file used to format it (cs65f12.sty). In addition, there is a bibliography style (cs65f12.bst) and sample bibliography file (cs65f12.bib). These files are all in the `cs65/labs/04-05/` directory.

7.2 The First Page

Center the title, author's name(s) and affiliation(s) across both columns. Do not use footnotes for affiliations. Use the two-column format only when you begin the abstract.

Title: Place the title centered at the top of the first page, in a 15-point bold font. A long title should be typed on two lines without a blank line intervening. Approximately, put the title at 1in from the top of the page, followed by a blank line, then the author's names(s), and the affiliation on the following line. Do not use only initials for given names (middle initials are allowed). The affiliation should contain the author's complete address, and an email address. Leave about 0.75in between the affiliation and the body of the first page.

Abstract: Type the abstract at the beginning of the first column. The width of the abstract text should be smaller than the width of the columns for the text in the body of the paper by about 0.25in on each side. Center the word **Abstract** in a 12 point bold font above the body of the abstract. The abstract should be a concise summary of the general thesis and conclusions of the paper. It should be no longer than 200 words.

Text: Begin typing the main body of the text immediately after the abstract, observing the two-column format as shown in the present document.

Indent when starting a new paragraph. For reasons of uniformity, use Adobe's **Times Roman** fonts, with 11 points for text and subsection headings, 12 points for section headings and 15 points for the title. If Times Roman is unavailable, use **Computer Modern Roman** (LaTeX's default; see section **??** above). Note that the latter is about 10% less dense than Adobe's Times Roman font.

7.3 Sections

Headings: Type and label section and subsection headings in the style shown on the present document. Use numbered sections (Arabic numerals) in order to facilitate cross references. Number subsections with the section number and the subsection number separated by a dot, in Arabic numerals. Do not number subsubsections.

Citations: Citations within the text appear in parentheses as (Harris, 1955) or, if the author's name appears in the text itself, as Harris (1955). Append lowercase letters to the year in cases of ambiguities. Treat double authors as in (Hafer and Weiss, 1974), but write as in (Hana et al., 2006) when more than two authors are involved. Collapse multiple citations as in (Harris, 1967; Déjean, 1998).

References: Gather the full set of references together under the heading **References**; place the section before any Appendices, unless they contain references. Arrange the references alphabetically by first author, rather than by order of occurrence in the text. Provide as complete a citation as possible, using a consistent format.

The LaTeX and BibTeX style files provided roughly fit the American Psychological Association format, allowing regular citations, short citations and multiple citations as described above.

Appendices: Appendices, if any, directly follow the text and the references (but see above). Letter them in sequence and provide an informative title: **Appendix A. Title of Appendix**.

Acknowledgement sections should go as a last section immediately before the references. Do not number the acknowledgement section.

7.4 Footnotes

Footnotes: Put footnotes at the bottom of the page. They may be numbered or referred to by asterisks or other symbols.[1] Footnotes should be separated from the text by a line.[2]

7.5 Graphics

Illustrations: Place figures, tables, and photographs in the paper near where they are first discussed, rather than at the end, if possible. Wide illustrations may run across both columns. Do not use color illustrations as they may reproduce poorly.

Captions: Provide a caption for every illustration; number each one sequentially in the form: "Figure 1. Caption of the Figure." "Table 1. Caption of the Table." Type the captions of the figures and tables below the body, using 11 point text.

References

Hervé Déjean. 1998. Morphemes as necessary concept for structures discovery from untagged corpora. In *Proceedings of the ACL-98 Workshop on New Methods in Language Processing and Computational Natural Language Learning*.

Margaret A. Hafer and Stephen F. Weiss. 1974. Word segmentation by letter success varieties. *Information Storage and Retrieval*, 10:371–385.

Jirka Hana, Anna Feldman, Luiz Amaral, and Chris Brew. 2006. Tagging Portuguese with a Spanish tagger. In *Proceedings of the EACL-06 Workshop on Cross-Language Knowledge Induction*.

Zellig Harris. 1955. From phoneme to morpheme. *Language*, 31:190–222.

Zellig Harris. 1967. Morpheme boundaries within words: Report on a computer test. In *Transformations and Discourse Analysis Papers*. Department of Linguistics, University of Pennsylvania.

[1] This is how a footnote should appear.

[2] Note the line separating the footnotes from the text.