# Sentiment Classification using Machine Learning Techniques

Pranjal Vashaspati
Institution1
Institution1 address
pranjal@mit.edu

Cathy Wu
Institution2
First line of institution2 address
cathywu@mit.edu

## Abstract

*We implement a series of classifiers (Naive Bayes, Maximum Entropy, and SVM) to distinguish positive and negative sentiment in critic and user reviews. We apply various processing methods, including negation tagging, part-of-speech tagging, and position tagging to achieve maximum accuracy. We test our classifiers on an external dataset to see how well they generalize. Finally, we use a majority-voting technique to combine classifiers and achieve accuracy of close to 90% in 3-fold cross-validation.*

## 1. Introduction

Sentiment analysis, broadly speaking, is the set of techniques that allows detection of emotional content in text. This has a variety of applications: it is commonly used by trading algorithms to process news articles, as well as by corporations to better respond to consumer service needs. Similar techniques can also be applied to other text analysis problems, like spam filtering.

## 2. Previous Work

We set out to replicate Pangs work from 2002 on using classical knowledge-free supervised machine learning techniques to perform sentiment classification. They used the machine learning methods (Naive Bayes, maximum entropy classification, and support vector machines), methods commonly used for topic classification, to explore the difference between and sentiment classification in documents. Pang cited a number of related works, but they mostly pertain to classifying documents on criteria weakly tied to sentiment or using knowledge-based sentiment classification methods. We used a similar dataset, as released by the authors, and did our best to use the same libraries and pre-processing techniques.

In addition to replicating Pangs work as closely as we could, we extended the work by exploring an additional dataset, additional preprocessing techniques, and combining classifiers. We tested how well classifiers trained on Pangs dataset extended to reviews in another domain. Although Pang limited many of his tests to use only the 16165 most common ngrams, advanced processors have lifted this computational constraint, and so we additionally tested on all ngrams. We use a newer parameter estimation algorithm called Limited-Memory Variable Metric (L-BFGS) for maximum entropy classification. Pang used the Improved Iterative Scaling method. We also implemented and tested the effect of term frequency-inverse document frequency (TF-IDF) on classification results.

### 2.1. Language

All manuscripts must be in English.

### 2.2. Dual submission

By submitting a manuscript to CVPR, the authors assert that it has not been previously published in substantially similar form. Furthermore, no paper which contains significant overlap with the contributions of this paper either has been or will be submitted during the CVPR 2011 review period to **either a journal** or any conference (including CVPR 2011) or any workshop (including CVPR2011 workshops) **Note that this is consistent with CVPR2010 but a strengthening from some previous CVPR policy**. Papers violating this condition will be rejected and a list of violating authors may be included in the proceedings.

If there are papers that may appear to the reviewers to violate this condition, then it is your responsibility to: (1) cite these papers (preserving anonymity as described in Section 1.6 below), (2) argue in the body of your paper why your CVPR paper is non-trivially different from these concurrent submissions, and (3) include anonymized versions of those papers in the supplemental material.

### 2.3. Paper length

CVPR papers may be between 6 pages and 8 pages, with a $100 per page added fee. Overlength papers will simply not be reviewed. This includes papers where the margins and formatting are deemed to have been significantly altered from those laid down by this style guide. Note that this LaTeX guide already sets figure captions and references in a smaller font. The reason such papers will not be reviewed is that there is no provision for supervised revisions of manuscripts. The reviewing process cannot determine the suitability of the paper for presentation in eight pages if it is reviewed in eleven. If you submit 8 for review expect to pay the added page charges for them.

### 2.4. The ruler

The LaTeX style defines a printed ruler which should be present in the version submitted for review. The ruler is provided in order that reviewers may comment on particular lines in the paper without circumlocution. If you are preparing a document using a non-LaTeX document preparation system, please arrange for an equivalent ruler to appear on the final output pages. The presence or absence of the ruler should not change the appearance of any other content on the page. The camera ready copy should not contain a ruler. (LaTeX users may uncomment the `\cvprfinalcopy` command in the document preamble.) Reviewers: note that the ruler measurements do not align well with lines in the paper — this turns out to be very difficult to do well when the paper contains many figures and equations, and, when done, looks ugly. Just use fractional references (e.g. this line is $095.5$), although in most cases one would expect that the approximate location will be adequate.

### 2.5. Mathematics

Please number all of your sections and displayed equations. It is important for readers to be able to refer to any particular equation. Just because you didn't refer to it in the text doesn't mean some future reader might not need to refer to it. It is cumbersome to have to use circumlocutions like "the equation second from the top of page 3 column 1". (Note that the ruler will not be present in the final copy, so is not an alternative to equation numbers). All authors will benefit from reading Mermin's description of how to write mathematics.

### 2.6. Blind review

Many authors misunderstand the concept of anonymizing for blind review. Blind review does not mean that one must remove citations to one's own work—in fact it is often impossible to review a paper unless the previous citations are known and available.

Blind review means that you do not use the words "my" or "our" when citing previous work. That is all. (But see below for techreports)

Saying "this builds on the work of Lucy Smith [1]" does not say that you are Lucy Smith, it says that you are building on her work. If you are Smith and Jones, do not say "as we show in [7]", say "as Smith and Jones show in [7]" and at the end of the paper, include reference 7 as you would any other cited work.

An example of a bad paper just asking to be rejected:

> An analysis of the frobnicatable foo filter.
>
> In this paper we present a performance analysis of our previous paper [1], and show it to be inferior to all previously known methods. Why the previous paper was accepted without this analysis is beyond me.
>
> [1] Removed for blind review

An example of an acceptable paper:

> An analysis of the frobnicatable foo filter.
>
> In this paper we present a performance analysis of the paper of Smith *et al*. [1], and show it to be inferior to all previously known methods. Why the previous paper was accepted without this analysis is beyond me.
>
> [1] Smith, L and Jones, C. "The frobnicatable foo filter, a fundamental contribution to human knowledge". Nature 381(12), 1-213.

If you are making a submission to another conference at the same time, which covers similar or overlapping material, you may need to refer to that submission in order to explain the differences, just as you would if you had previously published related work. In such cases, include the anonymized parallel submission [**?**] as additional material and cite it as

> [1] Authors. "The frobnicatable foo filter", F&G 2011 Submission ID 324, Supplied as additional material `fg324.pdf`.

Finally, you may feel you need to tell the reader that more details can be found elsewhere, and refer them to a technical report. For conference submissions, the paper must stand on its own, and not *require* the reviewer to go to a techreport for further details. Thus, you may say in the body of the paper "further details may be found in [**?**]". Then submit the techreport as additional material. Again, you may not assume the reviewers will read this material.

Sometimes your paper is about a problem which you tested using a tool which is widely known to be restricted to a single institution. For example, let's say it's 1969, you have solved a key problem on the Apollo lander, and you believe that the CVPR11 audience would like to hear about your solution. The work is a development of your celebrated 1968 paper entitled "Zero-g frobnication: How being the only people in the world with access to the Apollo

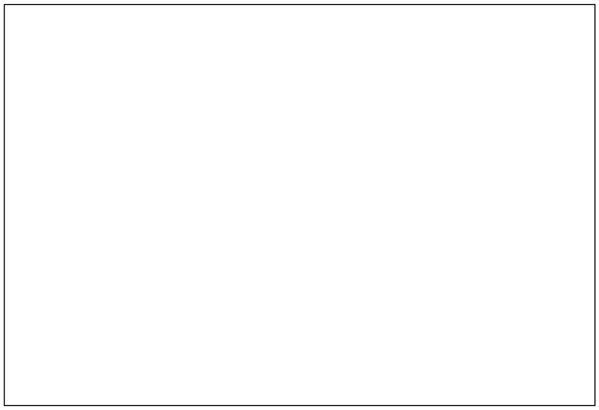

Figure 1. Example of caption. It is set in Roman so that mathematics (always set in Roman: $B \sin A = A \sin B$) may be included without an ugly clash.

lander source code makes us a wow at parties", by Zeus *et al*.

You can handle this paper like any other. Don't write "We show how to improve our previous work [Anonymous, 1968]. This time we tested the algorithm on a lunar lander [name of lander removed for blind review]". That would be silly, and would immediately identify the authors. Instead write the following:

> We describe a system for zero-g frobnication. This system is new because it handles the following cases: A, B. Previous systems [Zeus et al. 1968] didn't handle case B properly. Ours handles it by including a foo term in the bar integral.
>
> ...
>
> The proposed system was integrated with the Apollo lunar lander, and went all the way to the moon, don't you know. It displayed the following behaviours which show how well we solved cases A and B: ...

As you can see, the above text follows standard scientific convention, reads better than the first version, and does not explicitly name you as the authors. A reviewer might think it likely that the new paper was written by Zeus *et al*., but cannot make any decision based on that guess. He or she would have to be sure that no other authors could have been contracted to solve problem B.

FAQ: Are acknowledgements OK? No. Leave them for the final copy.

### 2.7. Miscellaneous

Compare the following:

| | |
|---|---|
| `$conf_a$` | $conf_a$ |
| `$\mathit{conf}_a$` | $\mathit{conf}_a$ |

See The TeXbook, p165.

The space after *e.g.*, meaning "for example", should not be a sentence-ending space. So *e.g.* is correct, *e.g.* is not. The provided `\eg` macro takes care of this.

When citing a multi-author paper, you may save space by using "et alia", shortened to "*et al.*" (not "*et. al.*" as "*et*" is a complete word.) However, use it only when there are three or more authors. Thus, the following is correct: " Frobnication has been trendy lately. It was introduced by Alpher [**?**], and subsequently developed by Alpher and Fotheringham-Smythe [**?**], and Alpher *et al*. [**?**]."

This is incorrect: "... subsequently developed by Alpher *et al*. [**?**] ..." because reference [**?**] has just two authors. If you use the `\etal` macro provided, then you need not worry about double periods when used at the end of a sentence as in Alpher *et al*.

For this citation style, keep multiple citations in numerical (not chronological) order, so prefer [**?**, **?**, **?**] to [**?**, **?**, **?**].

## 3. Formatting your paper

All text must be in a two-column format. The total allowable width of the text area is $6\frac{7}{8}$ inches (17.5 cm) wide by $8\frac{7}{8}$ inches (22.54 cm) high. Columns are to be $3\frac{1}{4}$ inches (8.25 cm) wide, with a $\frac{5}{16}$ inch (0.8 cm) space between them. The main title (on the first page) should begin 1.0 inch (2.54 cm) from the top edge of the page. The second and following pages should begin 1.0 inch (2.54 cm) from the top edge. On all pages, the bottom margin should be 1-1/8 inches (2.86 cm) from the bottom edge of the page for $8.5 \times 11$-inch paper; for A4 paper, approximately 1-5/8 inches (4.13 cm) from the bottom edge of the page.

### 3.1. Margins and page numbering

All printed material, including text, illustrations, and charts, must be kept within a print area 6-7/8 inches (17.5 cm) wide by 8-7/8 inches (22.54 cm) high.

### 3.2. Type-style and fonts

Wherever Times is specified, Times Roman may also be used. If neither is available on your word processor, please use the font closest in appearance to Times to which you have access.

MAIN TITLE. Center the title 1-3/8 inches (3.49 cm) from the top edge of the first page. The title should be in Times 14-point, boldface type. Capitalize the first letter of nouns, pronouns, verbs, adjectives, and adverbs; do not capitalize articles, coordinate conjunctions, or prepositions (unless the title begins with such a word). Leave two blank lines after the title.

AUTHOR NAME(s) and AFFILIATION(s) are to be centered beneath the title and printed in Times 12-point, non-boldface type. This information is to be followed by two blank lines.

Figure 2. Example of a short caption, which should be centered.

The ABSTRACT and MAIN TEXT are to be in a two-column format.

MAIN TEXT. Type main text in 10-point Times, single-spaced. Do NOT use double-spacing. All paragraphs should be indented 1 pica (approx. 1/6 inch or 0.422 cm). Make sure your text is fully justified—that is, flush left and flush right. Please do not place any additional blank lines between paragraphs.

Figure and table captions should be 9-point Roman type as in Figures 1 and 2. Short captions should be centred.

Callouts should be 9-point Helvetica, non-boldface type. Initially capitalize only the first word of section titles and first-, second-, and third-order headings.

FIRST-ORDER HEADINGS. (For example, **1. Introduction**) should be Times 12-point boldface, initially capitalized, flush left, with one blank line before.

SECOND-ORDER HEADINGS. (For example, **1.1. Database elements**) should be Times 11-point boldface, initially capitalized, flush left, with one blank line before, and one after. If you require a third-order heading (we discourage it), use 10-point Times, boldface, initially capitalized, flush left, preceded by one blank line, followed by a period and your text on the same line.

## 3.3. Footnotes

Please use footnotes[1] sparingly. Indeed, try to avoid footnotes altogether and include necessary peripheral observations in the text (within parentheses, if you prefer, as in this sentence). If you wish to use a footnote, place it at the bottom of the column on the page on which it is referenced. Use Times 8-point type, single-spaced.

| Method | Frobnability |
|---|---|
| Theirs | Frumpy |
| Yours | Frobbly |
| Ours | Makes one's heart Frob |

Table 1. Results. Ours is better.

## 3.4. Appendix A

## 3.5. Appendix B

## 3.6. References

List and number all bibliographical references in 9-point Times, single-spaced, at the end of your paper. When referenced in the text, enclose the citation number in square brackets, for example [**?**]. Where appropriate, include the name(s) of editors of referenced books.

## 3.7. Illustrations, graphs, and photographs

All graphics should be centered. Please ensure that any point you wish to make is resolvable in a printed copy of the paper. Resize fonts in figures to match the font in the body text, and choose line widths which render effectively in print. Many readers (and reviewers), even of an electronic copy, will choose to print your paper in order to read it. You cannot insist that they do otherwise, and therefore must not assume that they can zoom in to see tiny details on a graphic.

When placing figures in LaTeX, it's almost always best to use `\includegraphics`, and to specify the figure width as a multiple of the line width as in the example below

```
\usepackage[dvips]{graphicx} ...
\includegraphics[width=0.8\linewidth]
                {myfile.eps}
```

## 3.8. Color

Color is valuable, and will be visible to readers of the electronic copy. However ensure that, when printed on a

[1] This is what a footnote looks like. It often distracts the reader from the main flow of the argument.

| Test configurations | | | | Naive Bayes | | | MaxEnt | | | SVM | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Domain | Features | # of features | Frequency | + | - | ± | + | - | ± | + | - | ± |
| No-negation | Unigrams | 16165 | Frequency | 0.94 | 0.62 | 0.78 | - | - | - | 0.82 | 0.82 | 0.82 |
| No-negation | Unigrams | 16165 | Presence | 0.87 | 0.72 | 0.82 | 0.85 | 0.87 | 0.86 | 0.85 | 0.84 | 0.84 |
| No-negation | Bigrams | 16165 | Frequency | 0.92 | 0.64 | 0.78 | - | - | - | 0.77 | 0.81 | 0.79 |
| No-negation | Bigrams | 16165 | Presence | 0.89 | 0.73 | 0.81 | 0.79 | 0.82 | 0.81 | 0.8 | 0.81 | 0.8 |
| adjectives | Unigrams | 16165 | Frequency | 0.95 | 0.52 | 0.73 | - | - | - | 0.75 | 0.77 | 0.76 |
| default | Bigrams | 2633 | Frequency | 0.91 | 0.46 | 0.69 | - | - | - | 0.74 | 0.75 | 0.75 |
| default | Bigrams | 16165 | Frequency | 0.92 | 0.64 | 0.78 | - | - | - | 0.78 | 0.79 | 0.78 |
| default | Unigrams | 2633 | Frequency | 0.96 | 0.5 | 0.74 | - | - | - | 0.81 | 0.79 | 0.8 |
| default | Unigrams | 16165 | Frequency | 0.93 | 0.59 | 0.76 | - | - | - | 0.82 | 0.81 | 0.82 |
| default | Unigrams | maximum | Frequency | 0.95 | 0.49 | 0.72 | - | - | - | 0.82 | 0.81 | 0.82 |
| partofspeech | Bigrams | 16165 | Frequency | 0.96 | 0.47 | 0.71 | - | - | - | 0.82 | 0.82 | 0.82 |
| partofspeech | Unigrams | 16165 | Frequency | 0.96 | 0.54 | 0.75 | - | - | - | 0.82 | 0.81 | 0.81 |
| position | Bigrams | 16165 | Frequency | 0.96 | 0.49 | 0.73 | - | - | - | 0.77 | 0.78 | 0.78 |
| position | Unigrams | 16165 | Frequency | 0.93 | 0.58 | 0.76 | - | - | - | 0.81 | 0.82 | 0.82 |
| verbs | Unigrams | maximum | Frequency | 0.8 | 0.55 | 0.67 | - | - | - | 0.61 | 0.65 | 0.63 |
| adjectives | Unigrams | 16165 | Presence | 0.93 | 0.59 | 0.76 | 0.79 | 0.77 | 0.78 | 0.75 | 0.73 | 0.74 |
| default | Bigrams | 2633 | Presence | 0.86 | 0.64 | 0.75 | 0.75 | 0.75 | 0.75 | 0.73 | 0.75 | 0.74 |
| default | Bigrams | 16165 | Presence | 0.89 | 0.74 | 0.81 | 0.81 | 0.82 | 0.81 | 0.78 | 0.79 | 0.78 |
| default | Unigrams | 2633 | Presence | 0.84 | 0.8 | 0.82 | 0.84 | 0.82 | 0.83 | 0.78 | 0.82 | 0.8 |
| default | Unigrams | 16165 | Presence | 0.87 | 0.77 | 0.82 | 0.84 | 0.85 | 0.85 | 0.83 | 0.82 | 0.83 |
| default | Unigrams | maximum | Presence | 0.91 | 0.7 | 0.81 | 0.84 | 0.86 | 0.85 | 0.83 | 0.85 | 0.84 |
| partofspeech | Bigrams | 16165 | Presence | 0.89 | 0.73 | 0.81 | 0.84 | 0.84 | 0.84 | 0.79 | 0.82 | 0.8 |
| partofspeech | Unigrams | 16165 | Presence | 0.86 | 0.76 | 0.81 | 0.85 | 0.85 | 0.85 | 0.84 | 0.83 | 0.84 |
| position | Bigrams | 16165 | Presence | 0.87 | 0.66 | 0.76 | 0.82 | 0.83 | 0.82 | 0.73 | 0.76 | 0.74 |
| position | Unigrams | 16165 | Presence | 0.86 | 0.78 | 0.82 | 0.84 | 0.85 | 0.85 | 0.8 | 0.8 | 0.8 |
| verbs | Unigrams | maximum | Presence | 0.8 | 0.54 | 0.67 | 0.65 | 0.65 | 0.65 | 0.64 | 0.63 | 0.635 |
| adjectives | Unigrams | 16165 | TF-IDF | 0.82 | 0.6 | 0.71 | - | - | - | 0.79 | 0.76 | 0.77 |
| default | Bigrams | 2633 | TF-IDF | 0.92 | 0.46 | 0.69 | - | - | - | 0.76 | 0.71 | 0.74 |
| default | Bigrams | 16165 | TF-IDF | 0.9 | 0.68 | 0.79 | - | - | - | 0.83 | 0.74 | 0.79 |
| default | Unigrams | 2633 | TF-IDF | 0.85 | 0.52 | 0.74 | - | - | - | 0.81 | 0.79 | 0.8 |
| default | Unigrams | 16165 | TF-IDF | 0.88 | 0.68 | 0.78 | - | - | - | 0.83 | 0.77 | 0.8 |
| default | Unigrams | maximum | TF-IDF | 0.86 | 0.65 | 0.76 | - | - | - | 0.83 | 0.78 | 0.81 |
| partofspeech | Bigrams | 16165 | TF-IDF | 0.89 | 0.67 | 0.78 | - | - | - | 0.79 | 0.74 | 0.76 |
| partofspeech | Unigrams | 16165 | TF-IDF | 0.89 | 0.63 | 0.76 | - | - | - | 0.81 | 0.78 | 0.79 |
| position | Bigrams | 16165 | TF-IDF | 0.89 | 0.59 | 0.74 | - | - | - | 0.79 | 0.69 | 0.74 |
| position | Unigrams | 16165 | TF-IDF | 0.91 | 0.61 | 0.76 | - | - | - | 0.81 | 0.71 | 0.76 |
| verbs | Unigrams | maximum | TF-IDF | 0.64 | 0.57 | 0.6 | - | - | - | 0.62 | 0.66 | 0.64 |

monochrome printer, no important information is lost by the conversion to grayscale.

## 4. Final copy

You must include your signed IEEE copyright release form when you submit your finished paper. We MUST have this form before your paper can be published in the proceedings.

| Test configurations | | | | Naive Bayes | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Domain | Features | # of features | Frequency | ***** | **** | *** | ** | * | score |
| default | Unigrams | 16165 | Frequency | 0.72 | 0.68 | 0.53 | 0.34 | 0.24 | 0.74 |
| default | Unigrams | 16165 | Presence | 0.49 | 0.41 | 0.24 | 0.14 | 0.08 | 0.71 |
| default | Bigrams | 16165 | Presence | 0.50 | 0.42 | 0.26 | 0.13 | 0.10 | 0.70 |
| position | Unigrams | 16165 | Presence | 0.35 | 0.29 | 0.14 | 0.08 | 0.04 | 0.65 |
| partofspeech | Unigrams | 16165 | Presence | 0.45 | 0.37 | 0.20 | 0.11 | 0.06 | 0.69 |
| adjectives | Unigrams | 16165 | Presence | 0.76 | 0.73 | 0.61 | 0.45 | 0.36 | 0.70 |
| verbs | Unigrams | 16165 | Presence | 0.44 | 0.43 | 0.41 | 0.37 | 0.32 | 0.56 |
| default | Unigrams | maximum | Presence | 0.59 | 0.55 | 0.36 | 0.23 | 0.15 | 0.72 |
| position | Unigrams | maximum | Presence | 0.54 | 0.50 | 0.33 | 0.22 | 0.14 | 0.70 |
| partofspeech | Unigrams | maximum | Presence | 0.56 | 0.52 | 0.35 | 0.22 | 0.14 | 0.71 |
| adjectives | Unigrams | maximum | Presence | 0.76 | 0.73 | 0.61 | 0.45 | 0.36 | 0.70 |
| verbs | Unigrams | maximum | Presence | 0.44 | 0.43 | 0.41 | 0.37 | 0.32 | 0.56 |

| Test configurations | | | | MaxEnt | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Domain | Features | # of features | Frequency | ***** | **** | *** | ** | * | score |
| default | Unigrams | 16165 | Frequency | - | - | - | - | - | - |
| default | Unigrams | 16165 | Presence | 0.61 | 0.57 | 0.39 | 0.23 | 0.11 | 0.75 |
| default | Bigrams | 16165 | Presence | 0.63 | 0.59 | 0.45 | 0.28 | 0.26 | 0.68 |
| position | Unigrams | 16165 | Presence | 0.46 | 0.43 | 0.28 | 0.17 | 0.11 | 0.67 |
| partofspeech | Unigrams | 16165 | Presence | 0.55 | 0.50 | 0.32 | 0.20 | 0.10 | 0.72 |
| adjectives | Unigrams | 16165 | Presence | 0.75 | 0.72 | 0.62 | 0.45 | 0.37 | 0.69 |
| verbs | Unigrams | 16165 | Presence | 0.43 | 0.41 | 0.38 | 0.34 | 0.30 | 0.56 |
| default | Unigrams | maximum | Presence | 0.59 | 0.54 | 0.36 | 0.20 | 0.11 | 0.74 |
| position | Unigrams | maximum | Presence | 0.44 | 0.40 | 0.26 | 0.15 | 0.09 | 0.68 |
| partofspeech | Unigrams | maximum | Presence | 0.52 | 0.47 | 0.30 | 0.18 | 0.09 | 0.72 |
| adjectives | Unigrams | maximum | Presence | 0.75 | 0.72 | 0.62 | 0.45 | 0.37 | 0.69 |
| verbs | Unigrams | maximum | Presence | 0.43 | 0.41 | 0.38 | 0.34 | 0.30 | 0.56 |

| Test configurations | | | | SVM | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Domain | Features | # of features | Frequency | ***** | **** | *** | ** | * | score |
| default | Unigrams | 16165 | Frequency | 0.78 | 0.76 | 0.62 | 0.42 | 0.30 | 0.74 |
| default | Unigrams | 16165 | Presence | 0.58 | 0.54 | 0.38 | 0.25 | 0.14 | 0.72 |
| default | Bigrams | 16165 | Presence | 0.62 | 0.58 | 0.48 | 0.30 | 0.29 | 0.67 |
| position | Unigrams | 16165 | Presence | 0.42 | 0.39 | 0.27 | 0.39 | 0.42 | 0.50 |
| partofspeech | Unigrams | 16165 | Presence | 0.52 | 0.48 | 0.31 | 0.21 | 0.01 | 0.75 |
| adjectives | Unigrams | 16165 | Presence | 0.71 | 0.71 | 0.61 | 0.46 | 0.37 | 0.67 |
| verbs | Unigrams | 16165 | Presence | 0.45 | 0.45 | 0.42 | 0.38 | 0.32 | 0.57 |
| default | Unigrams | maximum | Presence | - | - | - | - | - | - |
| position | Unigrams | maximum | Presence | - | - | - | - | - | |
| partofspeech | Unigrams | maximum | Presence | - | - | - | - | - | - |
| adjectives | Unigrams | maximum | Presence | 0.71 | 0.71 | 0.61 | 0.46 | 0.37 | 0.67 |
| verbs | Unigrams | maximum | Presence | 0.45 | 0.45 | 0.42 | 0.38 | 0.32 | 0.57 |