Identifying Burns in Medical Device Reports

Lydia Gleaves, David Broniatowski, PhD.
Department of Engineering Management & Systems Engineering, The George Washington University
lpgleaves@gwu.edu

School of Engineering & Applied Science
THE GEORGE WASHINGTON UNIVERSITY

Background

The US Food and Drug Administration (FDA) receives thousands of medical device reports (MDRs) every year. [1] To aid postmarket surveillance, FDA analysts must determine whether these MDRs indicate a widespread device malfunction that may endanger public health, while filtering out noise. Analyst responses often differ given the same MDRs, creating uncertainty in how to proceed. This uncertainty may allow dangerous devices to remain on the market or lead to life-saving devices being suspended for unrelated issues. Machine learning may be used to augment the assessment process and flag reports containing injuries.

Objective

Build a classifier capable of identifying whether a report text conveys a burn injury.

Methods

1. **Extract text features from "complete" corpus.** The 1,505 most frequent n-grams (n = 2, 3) were taken from the complete corpus.
2. **Select the most valuable features.** A principal component analysis was performed to narrow down the possible features to the 91 components accounting for 80% of the data variance.
3. **Fit three classifiers to the data.** A logistic regression, Naive Bayes classifier, and support vector machine were fit to the data, tuning for high recall for reports containing burns (according to analyst consensus)..
4. **Assess classifier performance.** Each classifier was assessed using an analyst-annotated dataset of 499 MDRs, with the "correct" answer defined as the analyst consensus. Precision and recall across each category, overall accuracy, and average runtime were compared across the three classifiers to make a conclusion.

Contact Information

Lydia Gleaves — lpgleaves@gwu.edu
David Broniatowski — broniatowski@gwu.edu

Department of Eng. Mgmt. & Sys. Eng.
800 22nd St NW
Washington, DC 20052

Acknowledgements

We would like to thank Keith Walker for gathering the initial data.

Description of Data

This work relies on two sets of text containing the string "burn": a "complete" corpus of 64,000+ MDRs containing the string "burn" and a corpus of 499 MDRs annotated by 5 FDA analysts individually.

The FDA analysts were asked 1) Did the device appear to cause or contribute to a burn to a person? and 2) If so, was the burn life-threatening or permanently damaging? given the choices "Yes," "No," and "Not Enough Information" to answer each question. Inter-rater agreement was $\kappa = 0.65$ using Fleiss' kappa, which is "substantial agreement." [2] [3]

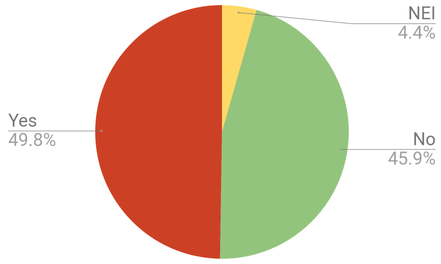

Figure 1. Distribution of answers to question #1

Feature Extraction and Selection

Text features were extracted by finding the 1,505 most common bi- and trigrams (n-grams, where n = 2, 3) in the complete corpus. These n-grams are often used in medical decision-making contexts [4].

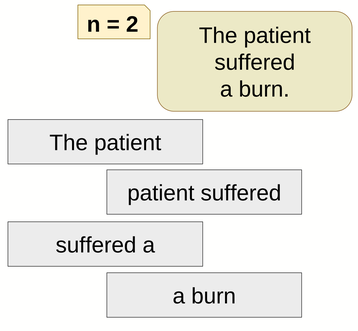

Figure 2. Bigrams extracted from sample sentence.

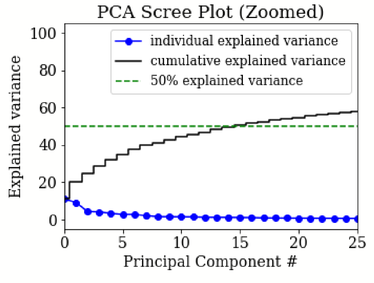

Figure 3. Scree plot of principal component analysis, zoomed to the 50% mark, demonstrating the usefulness of these components.

Text features were chosen for use in the classifiers by performing a principal component analysis and choosing those components which accounted for 80% of the total variance. A term-document matrix was created for the corpus, with each term being an n-gram feature and each document a device report, and the analysis was performed on the normalized data. This reduced the number of features from 1,505 to 91.

Results: Classifier Performance

	"Yes" Recall	10-Fold CV Accuracy	Average Run Time
Logistic Regression	0.96	0.73	0.2790s
Naive Bayes Classifier	0.73	0.63	0.0858s
Support Vector Machine	0.96	0.68	0.1721s

Figure 4. Table of limited classifier performance results.

While the Naive Bayes classifier outperforms the logistic regression and support vector machine in terms of speed, it underperforms in every other respect and can be discarded. The logistic regression has a higher average accuracy than the support vector machine, but most of the ~5% difference lays in incorrectly classifying reports as more severe than they are, which is vastly preferable to underrating report severity. With equal recall for true burns and a higher speed, the recommendation from this work is the support vector machine.

Conclusion: Recommend Support Vector Machine

The logistic regression had a higher average accuracy than the SVM, but most of the ~5% difference lays in incorrectly classifying "Not Enough Information" and "No" reports. With equal recall for true burns and a higher speed, the recommendation from this work for this purpose is the support vector machine.

Future Work: Assessment of "serious/life-threatening"

This classifier may not generalize to all reports if language surrounding other injuries is too dissimilar in structure (e.g., no clear signals like "degree burn"). Future work should explore this feature selection method with different types of MDRs. Additionally, this classifier uses only the text of the report to make a judgment. Human analysts may rely on other aspects of the report. Future work should endeavor to identify these other aspects and meaningfully quantify them if possible.

References

[1] FDA.
[2] Fleiss, J. L. (1971). Psychology Bulletin 76(5), 378-382.
[3] Landis, J. R., Koch, G. G. Biometrics (1977). 33(1), 159-174.
[4] Houvardas, J., Stamatatos, E. (2006). In: Euzenat, J., Domingue, J. (eds.) Artificial Intelligence: Methodology, Systems, and Applications, AIMSA, 4183.