UNITED STATES MILITARY ACADEMY

WEST POINT, NEW YORK

MA491 THESIS

USING MUTUAL INFORMATION FOR THE EFFICIENT CREATION OF DOMAIN SPECIFIC LEXICONS

by

CDT Michael Garrett

May 2018

Advisor: Captain Patrick Kuiper
Advisor: Major Karoline Hood

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE		Form Approved OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE May 2018	3. REPORT TYPE AND DATES COVERED

4. TITLE AND SUBTITLE Using Mutual Information for the Efficient Creation of Domain Specific Lexicons	5. FUNDING NUMBERS
6. AUTHOR(S) CDT Michael Garrett	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) United States Military Academy West Point, NY 10996	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) NA	10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

12a. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public release; distribution is unlimited	12b. DISTRIBUTION CODE

13. ABSTRACT (maximum 200 words)

We present a methodology to generate a domain specific lexicon to assess the population sentiment towards the US Military's disaster relief efforts immediately after Hurricane Maria in 2017. To establish a baseline sentiment, we leverage a corpus of geolocated Tweets from Twitter in Puerto Rico and employ the general purpose Multi-Perspective Questioning Answering (MPQA) lexicon and a common word polarity scoring method to infer the sentiment of each Tweet. To improve sentiment predictions, we build upon established literature which uses Point Wise Mutual Information (PMI) to assign word polarities. We suggest averaging PMI over a number of collocated words of known polarity, bootstrapping from the MPQA lexicon, building a measure of Mutual Information (MI). MI is used to build an improved domain specific MPQA lexicon, which is subsequently employed to evaluate Tweet sentiment and compared to the baseline predictions using the Puerto Rico corpus. Results indicate that updating the MPQA lexicon using the MI measure increases the accuracy of sentiment analysis, especially in extreme environments.

Key Words: Unsupervised Learning, Natural Language Processing, Lexicon, Mutual Information, Disaster Relief

14. SUBJECT TERMS			15. NUMBER OF PAGES
			16. PRICE CODE
17. SECURITY CLASSIFICATION OF REPORT Unclassified	18. SECURITY CLASSIFICATION OF THIS PAGE Unclassified	19. SECURITY CLASSIFICATION OF ABSTRACT Unclassified	20. LIMITATION OF ABSTRACT OF UL

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. 239-18

THIS PAGE INTENTIONALLY LEFT BLANK

USING MUTUAL INFORMATION FOR THE EFFICIENT CREATION OF DOMAIN SPECIFIC LEXICONS

CDT Michael Garrett
CDT, United States Army
B.S., United States Military Academy, 2018

Submitted in partial fulfillment of the
requirements for the degree of
BACHELOR OF SCIENCE
in **MATHEMATICAL SCIENCES**
from the
UNITED STATES MILITARY ACADEMY
May 2018

Author: CDT Michael Garrett

Advisory Team: Captain Patrick Kuiper
Instructor, Department of Mathematical Sciences

Major Karoline Hood
Assistant Professor, Department of Mathematical Sciences

Colonel Steve Horton
Chairman, Department of Mathematical Sciences

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

We present a methodology to generate a domain specific lexicon to assess the population sentiment towards the US Military's disaster relief efforts immediately after Hurricane Maria in 2017. To establish a baseline sentiment, we leverage a corpus of geolocated Tweets from Twitter in Puerto Rico and employ the general purpose Multi-Perspective Questioning Answering (MPQA) lexicon and a common word polarity scoring method to infer the sentiment of each Tweet. To improve sentiment predictions, we build upon established literature which uses Point Wise Mutual Information (PMI) to assign word polarities. We suggest averaging PMI over a number of collocated words of known polarity, bootstrapping from the MPQA lexicon, building a measure of Mutual Information (MI). MI is used to build an improved domain specific MPQA lexicon, which is subsequently employed to evaluate Tweet sentiment and compared to the baseline predictions using the Puerto Rico corpus. Results indicate that updating the MPQA lexicon using the MI measure increases the accuracy of sentiment analysis, especially in extreme environments.

Key Words: Unsupervised Learning, Natural Language Processing, Lexicon, Mutual Information, Disaster Relief

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF FIGURES

THIS PAGE INTENTIONALLY LEFT BLANK

ACKNOWLEDGMENTS

I would like to thank my two advisors, Major Hood and Captain Kuiper, for providing me the expertise and coding needed to spring board into the field of Natural Language Processing. Their assistance has helped spark a personal interest in this cross section of computer science, mathematics, and linguistics that I hope to further explore post-graduation.

THIS PAGE INTENTIONALLY LEFT BLANK

I. INTRODUCTION

We focus our analysis on social media as a means to gauge the feelings of a population. Specifically, we examine the Tweets from the population of Puerto Rico in the months following Hurricane Maria in 2017. Tweets provide a dynamic body of text that tests the flexibility of our Sentiment Analysis (SA) algorithm to assess a varied body of text filled with informal language. The Puerto Rico corpus contains Tweets relating to military disaster relief operations providing insight into how the polarity of words change in military context.

From our work we conclude that Tweets are conducive for SA, because they are brief and discreet. Additionally, we find that one can use Mutual Information to quickly create a domain specific lexicon which significantly boosts the accuracy of SA.

THIS PAGE INTENTIONALLY LEFT BLANK

II. BACKGROUND

A. NATURAL LANGUAGE PROCESSING

Natural Language Processing (NLP) is a field of study that leverages computers to interpret natural human language. Work in the field of NLP began soon after the advent of computers, however researchers quickly encountered problems concerning how to code the many nuances of language [3]. Our work attempts to provide a statistical method for sentiment analysis that avoids many of the linguistic problems encountered by early NLP researchers.

Mathematicians, computer scientist and linguist in the 1950s and 1960s worked side by side to break down human languages into a codeable form. One key step for NLP is the ability to break down, or parse, written language into word or phrase units. Parsing allows for word processors to identify things like sentence fragments, or run-on sentences, but does not remedy the issues posed by ambiguous words or words with multiple meanings. For example in the sentence "Jim sent a large package and shirt," is the size of the shirt large or only the package? Imagine if an automated request were built off this input, what size shirt should be ordered? One solution to this problem would be to encode a rule that assumes consistency across conjunctions words [4], but this can pose problems if absolute accuracy is needed, and to fully encompass the nuances of natural language one needs a burdensome number of rules. Machine learning techniques have led NLP researchers toward using statistics to infer meaning instead of using hard coded rules, as done by early researchers in NLP [5]. Machine learning uses statistical methods, like cluster analysis, to find patterns in a test sample, which researchers can then compare to results in pre-screened text. Our algorithm follows this trend to rely on statistics over linguistic rules.

B. SENTIMENT ANALYSIS

Within the past 20 years researchers have attempted to apply NLP to the activity of Sentiment Analysis. Sentiment Analysis, also referred to as *opinion mining*, attempts to use computers to understand if peoples' written opinions reflect positive or negative attitudes towards the subject [4]. Sentiment analysis has flourished with the advent of e-commerce, and social media. With e-commerce, businesses have near instantaneous access to large amounts of customer feedback in the form of product reviews. The large quantity of reviews

posted on sites like “Amazon”, and “Walmart,” make it impractical to have humans read and determine if the writer of each post feels negatively or positively toward the product. Like wise, social media posts can be used to judge the opinion of the masses toward everything from political decisions to new style trends. The computer algorithm substitute for human readers, uses statistical methods to determine whether the existence and order of words in an opinion constitute the opinion as negative or positive [4]. Key words for determining the sentiment of an opinion, like “love” and “hate”, are referred to as *sentiment words* or *subjectivity words*. Opinion words are often organized in *sentiment lexicons*, which assign a degree of positivity or negativity for each word. While there exists general lexicons, it has become common for researchers to create and use specific lexicons when exploring fields with unique vernaculars, like law or medicine [2]. Researchers in the field of medicine determined whether or not to add words dependent on how frequently unidentified words occurred[2]. We are interested in gauging the sentiment of Tweets following Hurricane Maria to determine if there is value in creating a military specific lexicon.

C. DOMAIN SPECIFIC LEXICON

Domain Specific Lexicons can help researchers using SA get more accurate results than if they used a general lexicon [1] . Imagine a scenario where researchers want to gauge only the sentiment of teenagers toward a product. A domain specific lexicon created to capture the more casual use of words by teenagers can help prevent researchers from overestimating or underestimating their sentiment. Like the fields of law and science, the military has a unique vernacular that could benefit from its own lexicon. Our research strives to find a an efficient method for making a military specific lexicon.

D. NOISY CHANNELS

The concept of a “noisy channel first appeared in the field of communication. A noisy channel is some medium that changes a messages content during transmission [6]. Similar to how irregularities in air space change a radio transmissions message, irregularities in natural language can mask the intended meaning of the speaker from the receiver. In our case the sender is a human writer and the receiver is a computer.

For any message transmitted across a noisy channel there is are two distributions involved for each given message. The first distribution X corresponds to the message before it is sent and the distribution Y corresponds to the message which is received on the other

side of the channel. If there is not channel interference, than these distributions would be identical, but this is rarely the case. In the transmission of the two bit binary number 1 and 0 there is some probability P(p) that each bit will be transmitted across the channel correctly, and a complementary probability P(1-p) that the bit flips.

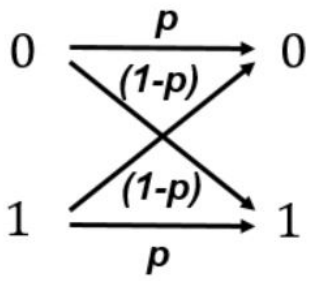

Figure 1. Noisy Channel Distribution

To eliminate the problems of noisy channels one must understand the characteristics of a channel. In the field of NLP these channel characteristics might take the form of syntax rules or the effects of particular words such as negators. With something as complex as natural language, which has millions of possible inputs and outputs, it is impossible to completely capture every relationship and achieve a zero error transmission, but we believe Mutual Information (MI) may allow us to infer relationships among words situated in close proximity to one another.

E. TURNEY ALGORITHM

In the early 2000s information technology researcher Peter Turney presented a simple method for determining sentiment in an article titled "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification or Review"[7]. Turney's method leverages Point Wise Information (PMI) to determine the polarity of unknown words.

$$PMI(X;Y) \; = \; log_2 \frac{p(x,\; y)}{p(x)p(y)}$$

PMI simply compares how often two objects, x and y, occur together versus occur apart. If x and y do not occur together than the PMI will be zero, while a high percentage of co-occurrence results in a large positive PMI.

PMI is of particular interest to our research, because of its simplicity. Turney's PMI method does not rely on hard coded semantic rules and therefore users may easily apply the method to varied contexts and languages. Our algorithm is largely based off of Turney's with the key difference being we use the average of many combinations of the target word and subjectivity words.

Turney's algorithm uses three inputs to calculate the sentiment of a word: a target word of unknown polarity, a positive or negative subjectivity, and a corpus. The corpus is the body of text through which the algorithm will search for the occurrence of the Target and subjectivity word.

Let **S** represent sentiment { **pos, neg** } word pair
Let **T** represent a list of target words for sentiment update
Let **C** represent a corpus of data for training
Initiate ***sentiment list*** = { }
→ For **t** in **T:**
→ Find freq(**t, S**[**pos**]), freq(**t, S**[**neg**]), freq(**S**[**pos**]), freq(**S**[**neg**]) in C
→ Calculate $\mathbf{PMI}_t = \log_2 \frac{\text{freq}(\mathbf{t, S[pos]}) * \text{freq}(\mathbf{S[neg]})}{\text{freq (t, S [neg]) * freq (S[pos])}}$
sentiment list[**t**] = $\mathbf{PMI}_t$
→ Return ***sentiment list***

Turney's algorithm provides much of our algorithm's mathematical foundation, but uses a very broad corpus and a small number of subjectivity words to assess the target word's polarity. Our algorithm narrows the corpus to give context to the results, and uses many subjectivity words to improve the result's accuracy.

III. METHODS

Our method for creating a military specific lexicon takes place in three steps. The first step required that we filter our corpus of 80 thousand Puerto Rico Tweets to 15 thousand exclusively English Tweets in order to match our exclusively English lexicon. Next we found the polarity of the most frequent words within the corpus by applying a variation on Turney's PMI algorithm to a corpus of 1000 randomly selected Tweets. Our set of target words include the most common subjectivity words found in the Tweets, as well as the subjectivity words included with MPQA. Using the most common words from the Twitter corpus helps make the updated lexicon domain specific.

We make two important changes to Turney's algorithm for the purpose of SA. First, instead of using only a pair of positive and negative sentiment words, our algorithm uses multiple combinations. Second, our algorithm averages the PMI from each iteration in an attempt to be more robust.

Let **S** represent a list of sentiment { **pos, neg** } word pairs
Let **T** represent a list of target words for sentiment update
Let **C** represent a corpus of data for training
Initiate ***sentiment list1*** = { }
Initiate ***sentiment list2*** = { }
→ For **t** in **T:**
    → For **s** in **S:**
        → Find freq(**t, s**[**pos**]), freq(**t, s**[**neg**]), freq(**s**[**pos**]), freq(**s**[**neg**]) in C
        → Calculate $\mathbf{PMI}_s = \log_2 \frac{\text{freq}(\mathbf{t, s}[\mathbf{pos}]) * \text{freq}(\mathbf{s}[\mathbf{neg}])}{\text{freq (t, s [neg]) * freq (s[pos])}}$
        → ***sentiment list1***[**s**] = **PMI**s
    → *sentiment list2*[**t**] = Σ *sentiment list1* [**s**] / Length S
→ Return ***sentiment list2***

After evaluating the polarity of a target word with our algorithm we edit the the general MPQA lexicon to create a domain specific lexicon, referred to as MPQA'. We determine whether or not to update a Target word's polarity by using the word's corresponding MI score. A word was considered positive if it has an MI greater than 0.10 and negative if the MI score was less than -0.10. In step two, we use the MPQA' and MPQA to perform

SA on the 1000 randomly selected Tweets. We determine the polarity of individual Tweets by measuring the ratio of positive and negative words within each Tweet. These same steps are then repeated using the MPQA'. In step three, we compare the results of each lexicon's SA, with the SA conducted by humans. Three native English speakers read each of the 1000 Tweets and determined whether they were positive, negative or neutral.

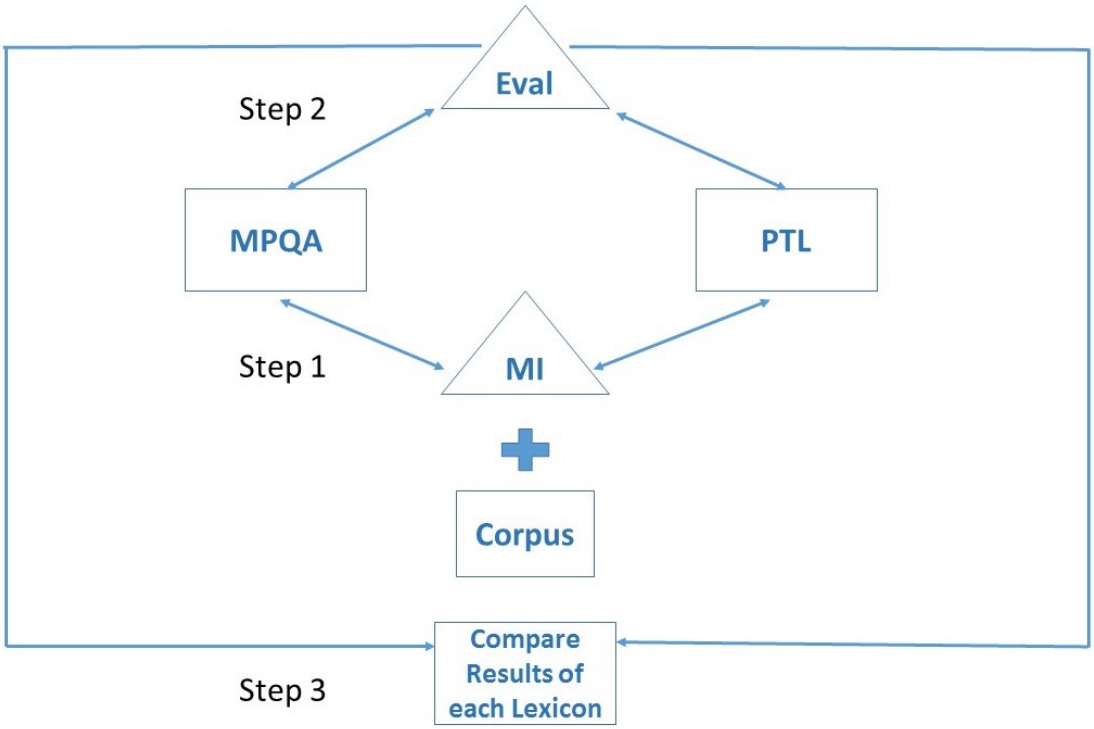

Figure 1. Method for Creating and Evaluating a Domain Specific Lexicon

IV. RESULTS

We assume that human judgment for sentiment polarity of a Tweet as the ground truth, so it is important to note that all three human readers categorized the sentiment of the 1000 Tweet test set similarly. When comparing the sentiment conclusions using MPQA' versus MPQA the results indicate that MPQA' identifies a significantly higher amount of negative Tweets. Referencing Figure 1, we observe that the amount of negative Tweets identified by MPQA' is a proportion much closer to the human validated results. The higher proportion of negative Tweets is intuitive given the context of natural disaster operations. These results indicate that using MPQA', created with the proposed algorithm, significantly improves the accuracy of sentiment analysis.

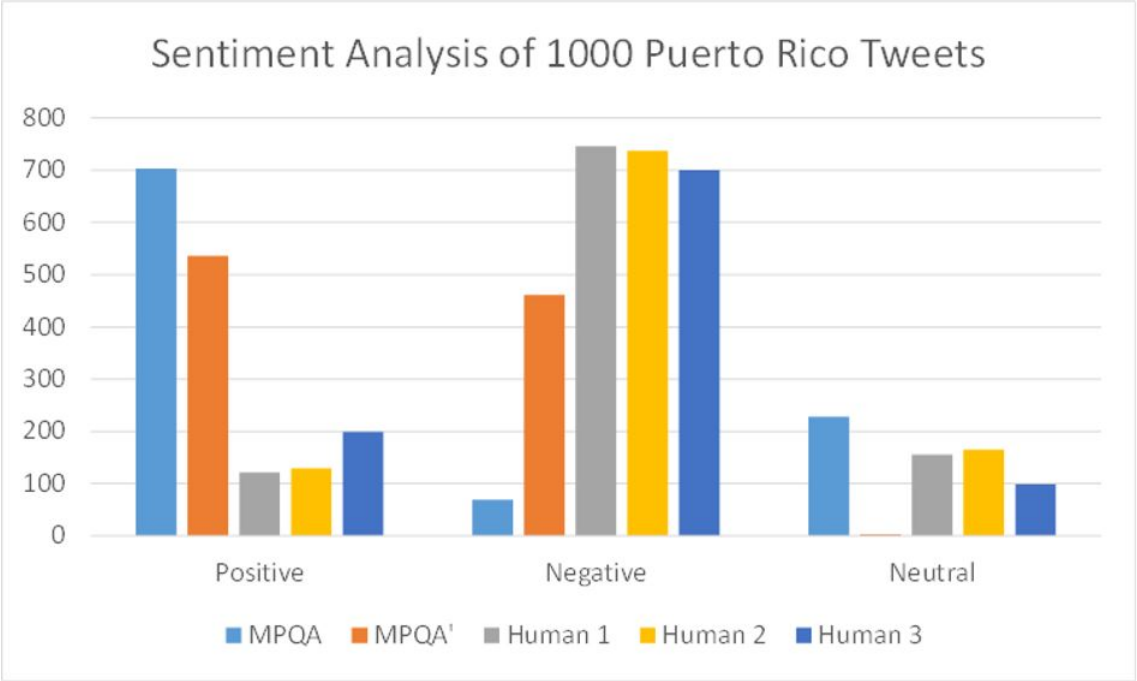

Figure 1. Sentiment Analysis results using MPQA and MPQA'

THIS PAGE INTENTIONALLY LEFT BLANK

V. ANALYSIS

The natural disaster context of the Tweets could explain the negative results of words that traditional lexicons attribute positive polarities. For example MPQA assigns "electricity" a neutral polarity, while our results assign a negative polarity.

This polarity difference presents a new set of challenges for updating a lexicon. It is difficult to know with only a single set of data if a drop in "electricity's" polarity is unique to Hurricane Maria or can be assumed for all disasters. We desire to confirm if a word's new polarity is unique, because if it true in only a single situation, then updating its polarity in a traditional lexicon, or adding a previously unlisted word, may hurt the lexicon's robustness.

THIS PAGE INTENTIONALLY LEFT BLANK

VI. CONCLUSION

The information that may be gained by autonomously assessing written natural language grows as people increasingly use forms of electronic communication like Social Media. We attempt to further progress in the field of NLP by providing a simple method for improving sentiment analysis accuracy given a specific domain, particularly in an extreme environment. The proposed method uses Mutual Information to assess the polarity of words, and then updates a general lexicon in order to create a domain specific lexicon.

Our results indicate that one can quickly create domain specific lexicons with MI. Our method's relative speed and flexibility make it a viable tool with which military commanders can attain a quick pulse on the population's feelings toward their units and the actions.

It is clear after looking at our variation of Turney's algorithm that creating the MI measure for each target word is an $O(3)$ operation. Further work is needed to optimize the calculation of this MI measure. Additionally, further investigation should be conducted to evaluate the criteria used to determine if the polarity of a word in a lexicon should be updated. Finally, a statistical analysis of SA results is needed to confirm whether this method of lexicon updating truly improves SA.

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF REFERENCES

[1] Yejin Choi and Claire Cardie. Adapting a polarity lexicon using integer linear programming for domain-specific sentiment classification. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2, pages 590–598. Association for Computational Linguistics, 2009.

[2] Carol Friedman. A broad-coverage natural language processing system. In Proceedings of the AMIA Symposium, page 270. American Medical Informatics Association, 2000.

[3] Karen Sparck Jones. Natural language processing: a historical review. In Current issues in computational linguistics: in honour of Don Walker, pages 3–16. Springer, 1994.

[4] Bing Liu. Sentiment analysis and opinion mining. Synthesis lectures on human language technologies, 5(1):1–167, 2012.

[5] Fabrizio Sebastiani. Machine learning in automated text categorization. ACM computing surveys (CSUR), 34(1):1–47, 2002.

[6] Claude Shannon. The zero error capacity of a noisy channel. IRE Transactions on Information Theory, 2(3):8–19, 1956.

[7] Peter D Turney. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In Proceedings of the 40th annual meeting on association for computational linguistics, pages 417–424. Association for Computational Linguistics, 2002.