Leveraging Mutual Information to Generate Domain Specific Lexicons *

Michael Garrett, Patrick Kuiper, Karoline Hood, and Dusty Turner

United States Military Academy, West Point, NY 10996
patrick.k.kuiper.mil@mail.mil

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.

Keywords: Unsupervised Learning · Natural Language Processing · Lexicon · Mutual Information · Disaster Relief.

1 Introduction

1.1 Problem Definition and Contribution

We focus our analysis on leveraging social media to gauge the feelings, or sentiment, of a large population towards an entity. Specifically, we examine Tweets from the population of Puerto Rico in the months following Hurricane Maria in 2017, in order to determine the resident populations' sentiment towards the US Military disaster relief efforts. Twitter delivers a dynamic corpus that provides valuable information to military and disaster relief decision makers, while challenging the flexibility of any Sentiment Analysis (SA) algorithm [3]. The Puerto Rico corpus contains approximately fifteen thousand English Tweets, presenting a varied body of text filled with informal language commenting on US Military disaster relief operations, among many other entities. The unique and extreme

* Supported by the Omar Bradley Fellowship.

environment associated with the generation of this corpus motivates the reasoning behind adjusting the polarity of words, given the context of the event.

From our work we conclude that a domain specific lexicon can be easily built using a simple Mutual Information (MI) algorithm, and that using this lexicon in sentiment analysis significantly improves accuracy.

1.2 Related Work

Natural Language Processing (NLP) Natural Language Processing (NLP) is a field of study leveraging computers to interpret human language. Work in the field of NLP began soon after the advent of computers; however, researchers quickly encountered problems concerning how to quantify the many nuances of languages [2]. We present an algorithm applied to a specific problem that could support any quantitative method within the NLP sub-field of sentiment analysis. The method avoids the linguistic rules associated with many NLP algorithms [3].

Shannon's Noisy Channel Coding Theorem and NLP Claude Shannon produced the first work considering the transmission of a message across a noisy channel, subsequently developing the field of Information Theory [4]. A noisy channel is a medium that communicates a message, which potentially changes content during signal transmission. Similar to how irregularities in the physical medium of air space change a radio transmission's message, irregularities in the perception of natural language, and the conceptual medium of language communication may perturb the intended message of the sender when transmitted to a receiver. We consider the sender to be a human writer with some internal sentiment they would like to communicate about an entity, while the receiver attempts to interpret the sentiment accurately given the sender's literal message.

There are two distributions associated with a noisy channel, $\mathcal{X}$ and $\mathcal{Y}$, which correspond to the distribution over the sender's messages transmitted and the distribution over the messages received. If there is no channel noise or interference, than these distributions are identical and completely dependent. For example, we may consider the transmission of a single bit binary message: 1 or 0. Over a noisy channel there is a probability p that each bit will be transmitted across the channel correctly, and a complementary probability $1 - p$ that the intended bit flips in error.

To transmit messages efficiently across a noisy channels it is critical to leverage the distributions of $\mathcal{X}$ and $\mathcal{Y}$. Foundational to Information Theory is the statistical measure of Mutual Information (MI), which quantifies the similarities between the two distributions and establishes the maximum capacity of a channel. When considering sentiment analysis in NLP, statistical characteristics may take the form of ambiguous grammar, complex vocabulary, or word dependencies. Even simple binary messages transmitted as natural language have almost limitless permutations for communication, rendering brute force enumeration of syntax rules and vocabulary algorithmically challenging. However, many state of

the art NLP sentiment analysis algorithms still attempt to leverage hard coded linguistic rules [3]. We assert that MI may provide a more efficient measure to infer sentiment.

Turney Algorithm The work of Church et. al. provided the first example of a statistical description of Lexicography or NLP using an Information Theoretic measure [1]. Church leveraged the measure of Mutual Information between two words to examine general association ratios for linguistic purposes not necessarily focused on sentiment. Turney extended this measure and presents a simple method for determining the sentiment of a specific target word [5]. Turney's method leverages Point-wise Mutual Information (PMI) to determine the polarity of a target word based on its collocation with a single pair of words of known polarity. This method is clearly well suited for lexicon generation.

$$PMI(\mathcal{X} = x, \mathcal{Y} = y) \;=\; log_2 \frac{p(x,\ y)}{p(x)p(y)} \tag{1}$$

When applied to analyzing text, PMI simply compares how often two words, $\mathcal{X} = x$ and $\mathcal{Y} = y$, occur together verses 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. Equation 1 shows the precise calculation of PMI.

Turney's use of PMI provides a method to generate a domain specific lexicon, which are an effective resource to improve sentiment analysis [3]. Turney's PMI method does not rely on hard coded semantic rules and therefore users may easily apply the method to varied contexts, languages, and algorithms.

2 Data and Methods

2.1 Puerto Rico Data Set

Hurricane Maria struck Puerto Rico with category four winds on 20 September 2017. The United States Military and disaster relief agencies responded shortly after this event. Upon arriving, military commanders requested a tool to quickly and accurately assess the sentiment of the population towards relief efforts. This analysis provides these disaster relief stakeholders with this capability.

We use Tweets as our corpus of data to assess the sentiment of Puerto Rico's population towards the US disaster relief efforts. The corpus is composed of over eighty-thousand Tweets that were scraped from Twitter and filtered for time and location to ensure they were written by members of the island's population following Hurricane Maria. We narrowed the data set to fifteen-thousand English text Tweets. This filter allowed for the use of the general purpose Multi-Perspective Question and Answer (MPQA) lexicon for assessing the polarity of undefined words [6].

2.2 Methodology

To find a baseline prediction of population sentiment using the Twitter corpus, we employed the established MPQA lexicon and a common sentiment classification algorithm which uses a ratio of positive to negative words to record (each Tweet's) polarity. This algorithm labels Tweets categorically either positive, negative, or neutral. With the time stamped polarities of thousands of Tweets, the average polarities of Tweets, binned by time, provides a tool to estimate population sentiment.

Initial evaluation of polarities assigned by the common ratio based sentiment algorithm proved to be inaccurate upon human validation. Further study indicated many words used from a negative perspective were considered positive by the MPQA lexicon. For example, Tweets with the word "Trump": referring to the President of the United States were overwhelmingly negative; however, the MPQA lexicon categorized "Trump" as a word indicating a positive perspective. The general purpose MPQA lexicon would have to be updated for the Puerto Rico domain specifically to provide accurate results for stakeholders.

To update MPQA based on a domain specific corpus we propose an algorithm closely related to Turney's with an important alteration: we calculate the average PMI of a target word with combinations of known polarity words, creating a measure of MI. The known polarity words are bootstrapped from the general purpose MPQA lexicon. The mathematical definition of MI is provided in Equation 2.

$$MI(\mathcal{X};\mathcal{Y}) \;=\; \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\; log_2\frac{p(x,\;y)}{p(x)p(y)} \tag{2}$$

We subsequently propose an update to Turney's algorithm, presented below in Figure 1.

Let **S** represent a list of all sentiment {**pos, neg**} word pair combinations in a Lexicon
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 $\mathbf{PMI}_s = \log_2 \frac{\text{freq(}\mathbf{t,\ s}[\mathbf{pos}])*\text{freq(}\mathbf{s}[\mathbf{neg}])}{\text{freq (}\mathbf{t,\ s}\text{ [}\mathbf{neg}\text{]) * freq (}\mathbf{s}[\mathbf{pos}])}$
→ ***sentiment list1***[**s**] = $\mathbf{PMI}_s$
→ ***sentiment list2***[**t**] = Σ ***sentiment list1*** [**s**] / Length S
→ Return ***sentiment list2***

Fig. 1. Modified Turney's Algorithm

To evaluate the proposed algorithm, we established a set of target words to be updated or added to the MPQA lexicon. We selected these words by determining the 100 most common words used in the corpus. With the list of target words to be updated, we applied the updated algorithm to assign new polarities. With the new polarities, we appended the MPQA lexicon, building a domain specific lexicon we call MPQA'.

Finally we randomly selected 1000 Tweets to act as a test set and executed sentiment analysis using each lexicon and the ratio based sentiment classifier. To assess performance between the two lexicons, we compared each lexicon's sentiment predictions to those labeled by three humans.

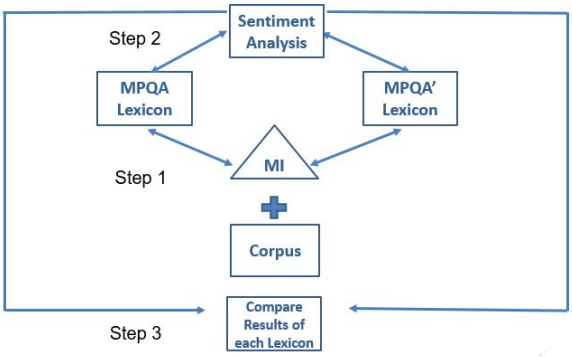

Fig. 2. Method for Creating and Evaluating a Domain Specific Lexicon

3 Results

We assume that human judgment for the sentiment polarity of a Tweet is the ground truth. 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 3, 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.

4 Conclusion and Future Work

The information that may be gained by autonomously assessing written natural language grows as people increasingly use forms of electronic communication

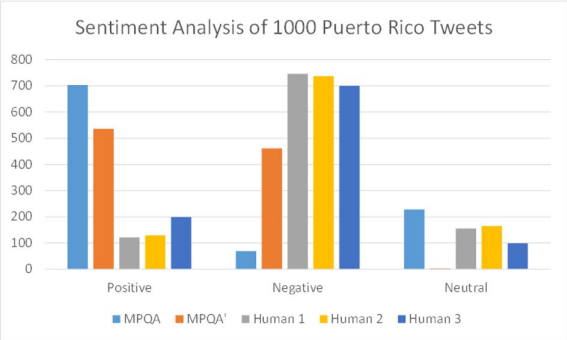

Fig. 3. Sentiment Analysis results using MPQA and MPQA'

such as Social Media. We attempt 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.

Referencing Figure 1, it is clear 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.

5 Acknowledgements

The authors thank the Department of Mathematical Sciences, United States Military Academy, and the General Omar Bradely Fellowship for support given throughout this research process.

References

1. Church, K.W., Hanks, P.: Word association norms, mutual information, and lexicography. Comput. Linguist. **16**(1), 22–29 (Mar 1990), http://dl.acm.org/citation.cfm?id=89086.89095
2. Jones, K.S.: Natural language processing: a historical review. In: Current issues in computational linguistics: in honour of Don Walker, pp. 3–16. Springer (1994)
3. Liu, B.: Sentiment analysis and opinion mining. Synthesis lectures on human language technologies **5**(1), 1–167 (2012)
4. Shannon, C.: The zero error capacity of a noisy channel. IRE Transactions on Information Theory **2**(3), 8–19 (1956)
5. Turney, P.D.: 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. pp. 417–424. Association for Computational Linguistics (2002)

6. Wilson, T., Wiebe, J., Hoffmann, P.: Recognizing contextual polarity in phrase-level sentiment analysis. In: Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing. pp. 347–354. HLT '05, Association for Computational Linguistics, Stroudsburg, PA, USA (2005). https://doi.org/10.3115/1220575.1220619, https://doi.org/10.3115/1220575.1220619