A Natural Language Normalization Approach to Enhance Social Media Text Reasoning

Long Hoang Nguyen, Andrew Salopek
Department of Computer Science
Texas Tech University
Lubbock, Texas
{*long.nguyen, dylan.salopek*}@*ttu.edu*

Liang Zhao
Information Science & Technology
George Mason University
Fairfax, Virginia
lzhao9@gmu.edu

Fang Jin
Computer Science Department
Texas Tech University
Lubbock, Texas
fang.jin@ttu.edu

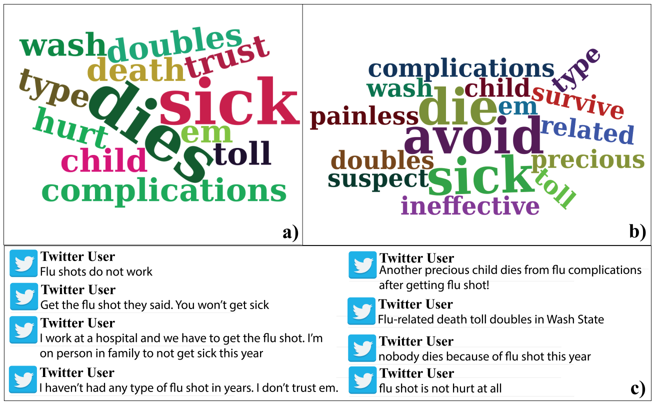
***Abstract*—Social media has become a popular data source to track and analyze societal events. Targeted domains such as election, civil unrest, and spreading disease all require a natural language normalization tool capable of extracting information pertinent to these domains accurately. Due to the unstructured language, short-length messages, casual posting styles, and homonyms, it is technically difficult and labor-intensive to remove barriers that may lead to inaccurate analysis. Because the fact that typos or other symbolic representations of sentiment may lead to lower frequency of term appearance, language preprocessing becomes critical and necessary to improve social media text reasoning.**

We propose a novel unsupervised preprocessing approach to enhance text understanding quality and illustrate this approach using one specific domain, flu shot reasoning. The proposed approach relies on a database of synonyms and opposite words and an algorithm to transform negative sentences into its affirmative form. In this form, the features and opinions are reflected accurately via transforming parts of speech. For instance, features are presented as nouns and opinions are presented as verbs or adjectives. The algorithm also corrects words if they are not correctly written and normalizes them to increase its frequency of appearance. The effectiveness of our algorithm is evaluated on the tweets dataset to answer why people are reluctant to take flu shots.

***Keywords*-Language Preprocessing; Information Retrieval; Sentiment Analysis; Social Media Reasoning**

I. Introduction

With the growth of internet and mobile devices, social media platforms such as Twitter and Facebook are experiencing an explosive level of growth. People nowadays get engaged with internet more frequently as a means for exchanging opinions, learning knowledge, and discussing societal events. Millions of worldwide social media accounts broadcast their daily observations on an enormous variety of domains, e.g., safety, politics [1], and disasters [2]. Publishing posts, giving comments, and sharing experiments are typical ways that people interact on online networks. With its gaining popularity, social media provides not just a platform for communication, but also a crucial platform for ongoing discussions of the latest news, and thus could serve as a societal sensor with which to track people's reactions to events [3], [4]. Twitter, for example, as one of the popular social network platforms, which allows people to share their thought by a tweet within 140 character length, has become a popular data source for monitoring and analyzing events. Researchers have utilized social networks as a means to analyze and figure out answers for different research questions. Using consumer product reviews, researchers can grab reasons to answer why some products gain better sale over other products and thus build a recommendation system by analyzing feedback comments [5]. Growth of varied social media has enabled economists to incorporate real-time indicators such as public emotion, anticipations and behaviors factors which possibly influence the market into modeling [6], [7].



Twitter User
Flu shots do not work

Twitter User
Get the flu shot they said. You won't get sick

Twitter User
I work at a hospital and we have to get the flu shot. I'm on person in family to not get sick this year

Twitter User
I haven't had any type of flu shot in years. I don't trust em.

Twitter User
Another precious child dies from flu complications after getting flu shot!

Twitter User
Flu-related death toll doubles in Wash State

Twitter User
nobody dies because of flu shot this year

Twitter User
flu shot is not hurt at all

Figure 1. Tweets transformation diagram. The left side a) shows the word cloud visualization without the proposed preprocessing methods. The right side b) shows the word cloud visualization with the proposed preprocessing methods. The bottom c) shows the original users' tweets used in the left word cloud.

Although information retrieval from plain text has been well studied [8], [9], analyzing tweets to reveal reasoning information requires more sophisticated techniques. Tweets are written in an unstructured language and often contain typos, non-standard acronyms, and mutual meanings. It may be in various forms of language presentations such as negative, affirmative or sarcasm, which makes the textual preprocessing phase become more important, in order to accurately catch the meaning of the sentence. In the text mining area, researchers rely heavily on the term frequency-inverse document frequency (TF-IDF) measure to classify tweets and extract features as well as opinions via lexicon analysis. Usually, features are represented as nouns and opinions are verbs or adjectives in the sentences [10]. For example, if a tweet saying *"nobody dies because of flu shot this year"*, then with traditional natural language processing, the word *"die"*, which is a verb will serve as the most important term and represent the opinion in this sentence. It may lead to misinterpretation of the text. Another example of a tweet in a negative form is *"flu shot does not hurt at all."*. The opinion word, which is representative for this statement, is *"hurt"*. While the other semantically similar tweet *"flu shot is painless"* has its extracted opinion *"painless"*, which is completely opposite to the other one, even though they express the same meaning. However, the traditional NLP processing makes it hard to treat them equal. Therefore, the term frequency in the bag of words model [11] may not be able to identify these correlations and thus result in lower accuracy of some critical information extraction. Moreover, as the tweets may be keyed in via a mobile, or in a quick sharing moment, or by a person who has some limitation in the language (foreign language, use short writing, etc), typos are a normal situation in social media platforms. This will lead to words getting unrecognized in a proper dictionary. All these features of Twitter data pose a challenge for reasoning methods developed for traditional media.

To overcome the above issues, we propose a solution in preprocessing phase that utilized the natural language processing approach. We normalize sentences to its affirmative form and translate words to its common synonym in order to increase term frequency in the corpus. In addition, any typo is parsed through a word correction mechanism to ensure none of the important words to be ignored. In particular, as shown in Figure 1, the left tweets will be transformed to *"everyone survives because of flu shot this year", "flu shot is painless"*. So the important terms got extracted will be *"survive"* (with one occurrence) and *"painless"* (with two occurrences). To avoid unnecessary new term occurrence which may dilute important features, any synonym or opposite word will be selected based on algorithms that prioritize choosing words from tweets under analysis rather than picking the terms from the synonyms and opposite database. As a result, the popular terms which are in the corpus will be clearly stood out from other terms, hence to increases the accuracy of reasoning task in later phases. In order to evaluate the effectiveness of the proposed approach, we conduct experiments on tweet messages related to flu shot and analyze reasons why people are reluctant to flu shots via user's sentiment analysis.

The main contribution of this paper includes:

- We propose a novel unsupervised preprocessing approach to enhance text understanding and illustrate this approach via flu shot reasoning.
- By building a database with synonyms and antonyms, we transfer various negative sentences into its affirmative format, while incorporate typo corrections at the same time, to reduce misinterpretation for language preprocessing.
- This smart normalization approach incorporates considerations of parts of speech for features and opinions, for instance, features are presented as nouns and opinions are transformed as verbs or adjectives, which can enhance text reasoning extensively.

The rest of this paper is organized as follow: Section II introduces related work, and section III presents the system framework and detailed algorithms. Section IV explains experimental results on the flu shot tweets and also results analysis. The conclusion and direction for future work are discussed in Section V.

II. Related Work

The related work falls into two categories.

A. Feature Extraction

Natural Language Processing (NLP) for decision making has been well studied using feature extraction and analysis methods [12]. Early algorithms used hard rules to parse and label grammar and phrases within input, using parse trees to derive the parts of speech of a string [13]. As NLP evolved, statistics and probability in NLP became more prominent [14]. Statistical natural language processing (SNLP) makes inferences of input data to improve rules which suggests that parsers should be designed based on sentences and their structure. Collobert and Weston [15] presented a neural network architecture for NLP allowing parsing of huge databases. The architecture was able to perform very well without syntactic features, proving that syntax is not a mandatory feature for semantic structure building and analyzing. Collobert et al. [16] proposed another neural network for NLP that is not task-specific, meaning the parser uses large amount of unlabelled data to let their training algorithm to discover and learn what the data represents and in turn provide a universal NLP for any data set.

B. Sentiment Analysis

Sentiment analysis and opinion mining through classification has become more popular since the mid-2000's, and is a widening area of research. More powerful machine learning tools are emerging, providing more insight to previously useless data. The bag-of-features used by Pang [17] training the models Naïve Bayes (NB) [18], Maximum Entropy (ME) [19] and Support Vector Machines (SVM) [20] were proven useful only for topic-dependent sentiment analysis and only performed well within the respective domain of analysis. For example, the study provided an accurate analysis of movie reviews and tried to reproduce the accuracy in other domains such as automobiles but were unable to do so. Read [21] began to analyze different type of sentiment, using domains, topics, temporal and language style in order to provide a more broad classifier to learn with and provided more accuracy across analyses. The study proposed the idea that classifiers may not be learning sentiment towards nouns, but rather learning semantic sentiment of associated words of those nouns [22]. Abbasi, Chen, and Salem [23] analyzed sentiment for violence and hate groups on web forums, gathering and analyzing user sentiment towards certain topics involving US and Middle East supremacy groups in both English and Arabic. They developed the entropy weighted genetic algorithm (EWGA) to identify more accurate features to be used in sentiment classes. The algorithm also delivered greater insight into writing style between to two groups. Pak and Paroubek [24] presented the ability to analyze Twitter data for positive, negative, and neutral sentiment based on tokens (emotion icons), such as “:-)”,“:)”,“:-(”,“:(” in order to determine the user’s sentiment. The classifier is based on the Naïve Bayes classifier, and features to derive a syntactic tree that shows a user’s emotion or statement of fact.

III. Framework

A. The Framework Overview

The social media text preprocessing framework is mainly composed of two parts, one is traditional preprocessing and the other is normalization, as shown in Figure 2. With social media data such as tweets as input, the preprocessing step will begin with our natural language normalization steps. Then it will perform traditional preprocessing steps such as stop words removal and stemming. In this study, we choose flu shot related tweets for analysis and reasoning. We will discuss the traditional processing steps first then the proposed normalization method afterwards. The detailed processing steps are illustrated below.

Traditional Preprocessing: Like any text analysis, classical preprocessing is necessary which help clean the tweets for the first step.

- *Filtering*: The tweets are filtered by some keywords such as *”flu” or ”flu shot”* to make it more relevant to our research domain.
- *Removal of special characters*: All special characters such as none-english characters or UTF-8 encoded characters are removed to reduce the amount of data to be processed.
- *Removal of stopwords*: There are some commonly used words, even though they are a part of the sentence but do not contribute much information are called stopwords, such as articles *“a“, “the“, “an“, “in“ or “for“* or the filtered key words *“flu“ or “flu shot“*, which are already appeared in all tweets.

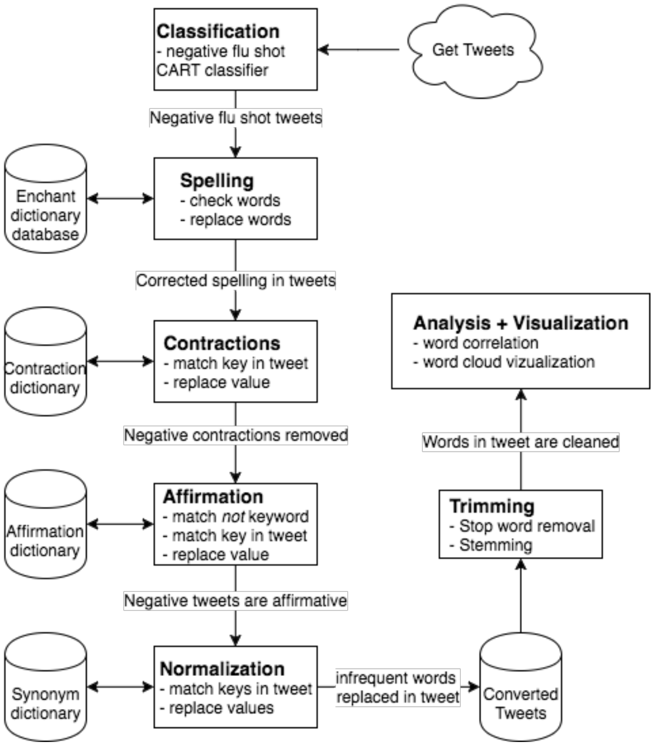


Figure 2. System Architecture

- *Lemmatization and Stemming*: The goal of these processes is to reduce inflectional forms with a morphological suffice like converting the words *“looking“, “looked“* and *“looks“* to its base form which is *“look“*, or chop off the endings and accept certain mistakes in word meaning. Stemming often includes removal of derivational affixes.

Language Normalization: Besides the traditional preprocessing, we added some additional steps to the traditional NLP model with further analysis at the sentence level using semantic similarity to enhance the weight for the identified words. The additional processes are presented as follows.

- *Spelling Correction*: Social media texts have a much higher probability of typos than other text format, due to its informal chat style or sticky keyboards. Thus it leads to unrecognized terms which can largely reduce some important terms’ frequencies. Therefore, in our framework, all tweets are parsed through a spelling corrector in order to fix any typo found in the tweets. The algorithm of this correction is described in detail in Section III-B.
- *Negative Contraction Transformation*: Many phrases are combined into contractions to shorten the words. Negative contractions such as *can’t, won’t, haven’t* will be separated into the long-form equivalents, *can not, will not, have not*. By separating these contractions into their long forms, it makes parsing and converting the

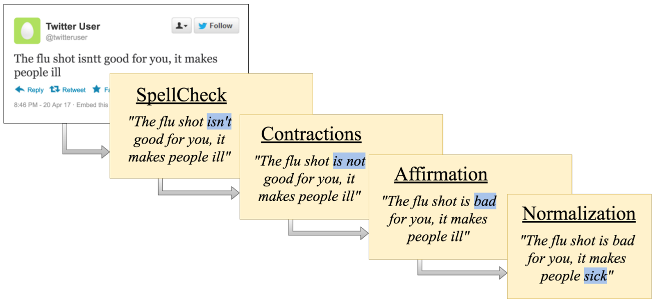

Twitter User
@twitteruser
Follow
The flu shot isntt good for you, it makes people ill
Reply
Retweet
8:46 PM - 20 Apr 17 · Embed this
SpellCheck
"The flu shot isn't good for you, it makes people ill"
Contractions
"The flu shot is not good for you, it makes people ill"
Affirmation
"The flu shot is bad for you, it makes people ill"
Normalization
"The flu shot is bad for you, it makes people sick"

Figure 3. Flowchart of data preprocessing

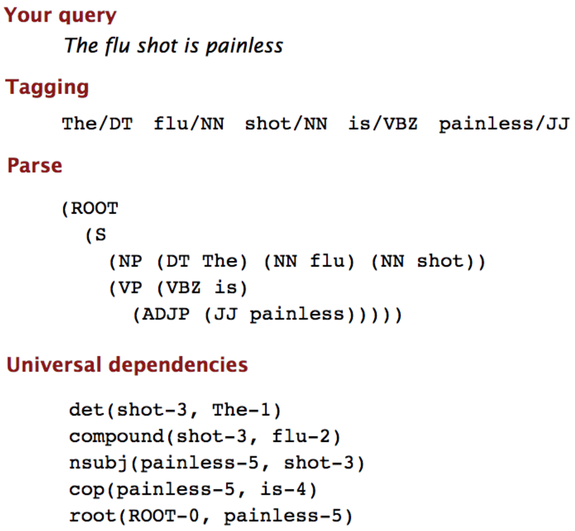
other parts of speech easier. Details are described in Section III-D1

- *Affirmative Subject Transformation*: As the subject of a sentence may be in its negative form such as *"nobody"*, *"noone"* etc. Hence there is no interest in the subject and it could lead to use of verb or adjective differently. This process will convert negative subject into its affirmative form and proceed with negative sentence transformation. Details are described in Section III-D2.
- *Affirmative Sentence Transformation*: Section III-D3 explains how to convert a negative sentence to its affirmative form. When a sentence is in its negative form, for example *i don't like flu shot*, the verbs or adjectives do not reflect actual opinion. In other words, verbs and adjectives are not direct attributes of the subject. By converting these sentences into its affirmative form, we can make the verbs and adjectives directly express opinions on the subject.
- *Word Normalization*: Once sentences are in affirmative forms, the algorithm increased frequency of a term by replacing its existing synonyms in the corpus to itself. This makes the original text hard to read, but can increase its weight which is beneficial for computational analysis. We called this process word normalization as it cleans data and converts various disordered terms to one standard criteria. Section III-E describes this process in detail.

Figure 3 describes a sample of the data flow for the normalization process. Given a tweet message, it will go through a spelling check, and then transform negative contractions, convert negative sentences to its affirmative form, and finally replace some terms to their synonyms.

B. Typos correction

Typos correction is to ensure getting the most accurate data for analyzing. A spell-checking library is available from PyEnchant that is built on top AbiWord's Enchant [25]. This library accesses 8 different dictionaries available on the web. Generally, the more dictionaries incorporated, the higher quality of the results returned. Although spell checking is a very significant and important part of tweets preprocessing, it actually comes after contraction transformations due to the fact that the spelling checker does not handle symbols such as apostrophes very well/accurately. The function separates the tweet into individual words and iterates over them, checking each word's spelling sequentially. If the word is misspelled, the function looks for suggestions, and returns the most probable one, as described in Algorithm 1.



```
Your query
        The flu shot is painless

Tagging
        The/DT  flu/NN  shot/NN  is/VBZ  painless/JJ

Parse
        (ROOT
          (S
            (NP (DT The) (NN flu) (NN shot))
            (VP (VBZ is)
              (ADJP (JJ painless)))))

Universal dependencies
        det(shot-3, The-1)
        compound(shot-3, flu-2)
        nsubj(painless-5, shot-3)
        cop(painless-5, is-4)
        root(ROOT-0, painless-5)
```

Figure 4. Example of Stanford NLP output from string of "The flu shot is painless".

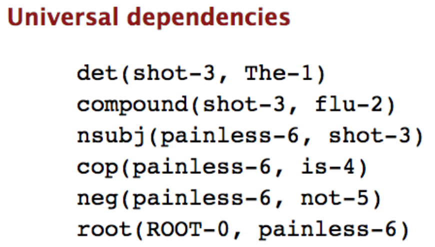
C. Language parser incorporation

The Stanford Parser [26], [27] generates the grammatical structure of an input and determines parts of speech (POS) of the sentence. From these parts of speech, a subject (feature) may be determined along with opinions (verbs and adjectives). If a sentence contains a negative word or identifier, the word and its dependency are listed as well. The parser recognizes negative phrases, looking for keyword such as *no* and *not*. Because of this, determining which tweets to parse and affirm becomes easier and more efficient. In Figure 4, the parse tree for the example string derives dependencies, showing that *shot* is the feature and *painless* is the adjective describing the *shot*. This is useful to help train the machine to analyze future tweets. In figure 5, the example shows that *shot* is the feature and *painless* is still the adjective describing the *shot* along with a negative dependency *not*.

D. Negative to affirmative transformation

Algorithm for the following transformations is described in Algorithm 1.

1) Negative contraction transformation: Negative contractions are handled by the Stanford NLP, identifying negative word such as *isn't, won't, haven't, etc*. This procedure is important because negative contractions need to be broken



Universal dependencies

```
det(shot-3, The-1)
compound(shot-3, flu-2)
nsubj(painless-6, shot-3)
cop(painless-6, is-4)
neg(painless-6, not-5)
root(ROOT-0, painless-6)
```

Figure 5. Example of negated string Stanford NLP output from string “The flu shot is not painless”

apart to derive the verb and the negation in order to correctly parse the data.

2) Negative subject transformation: To avoid negative subjects using verbs and adjectives incorrectly, we transformed them to their affirmative equivalent. For example, we would transform the tweet *Nobody enjoys getting the flu shot* to *Everybody does not enjoy getting the flu shot*. After this transformation, the tweet may be grammatically incorrect but is easier for the machine to understand.

3) Negative verb and adjective transformation: Negative verbs and adjectives can affect the analysis of the tweets by using false keywords. For example, the above tweet, *Everybody does not enjoy getting the flu shot* may be interpreted wrong by the machine learning program because of the keywords *enjoy* and *flu shot*. To avoid this potential problem, the program replaces negative verbs and adjectives with affirmative synonyms. The Stanford NLP parses the tweet data for any negative keywords or phrases, and determines the parts of speech that the following word has. If the following word is a verb or an adjective, it is then checked against a dictionary of antonyms. If it matches, the word is replaced by the new word, and the negative keyword is removed, ultimately affirming the sentence. In the example tweet, the program would find the phrase *not enjoy*, and replace it with *dislike*. The transformation gives a better input for the learning program to analyze.

E. Word normalization

We all know English may describe one thing in many different ways. For example, a Twitter user may post *Getting a flu shot is unwise*, or say *Getting a flu shot is foolish*, and even more other synonyms for the word “unwise”. In order to boost the TF-IDF for social media text analysis, it is necessary to convert a word and all of its synonyms to a standard word. In this case, you could use the words *unwise, stupid, foolish, idiotic or dumb* to describe disdain for a flu shot. When parsing through tweets and transforming them, we propose an algorithm to compare all the words in a tweet to a dictionary that contains synonyms and the most common word between them. This algorithm replaces a word with a more frequently used term in order to give a better analysis of an entire dataset.

Algorithm 1: Tweet preprocessing algorithm, includes spell checking, NLP, contraction handling, affirmation, and normalization

```
for tweet in tweet set do
    remove non-alphanumeric characters
    check word in tweet
    if word not in spelling dictionary then
        suggest and replace new word
        check next word
    else
        check next word
    end
    parse tweet with Stanford NLP
    if tweet contains negative then
        replace adj/verb with affirmative dictionary
          value
        remove negative word
    end
    normalize feature with the most frequent synonym
end
```

IV. EXPERIMENTAL RESULTS

A. CART Model for Tweet classification

We employed CART model for tweets classification to determine those who are reluctant to flu shots. CART (classification and regression tree) model is introduced by Leo Breiman in 1984. The technique is used to predict or classify the value of a target known as dependent variable based on the values of several inputs so called independent variables. *“A decision tree or a classification tree is a tree in which each internal (non-leaf) node is labeled with an input feature. The arcs coming from a node labeled with an input feature are labeled with each of the possible values of the target or output feature or the arc leads to a subordinate decision node on a different input feature. Each leaf of the tree is labeled with a class or a probability distribution over the classes“* [28]. In CART model, we use the degree of impurity of child nodes as a basis to select the best split in building the classification tree. The smaller the degree of impurity, the more skewed the class distribution. The impurity measures include:

- *Entropy*:
$$Entropy(t) = -\sum_{i=0}^{c-1} p(i|t) * log_2 p(i|t)$$
- *Gini index*:
$$Gini(t) = 1 - \sum_{i=0}^{c-1} [p(i|t)]^2$$
- *Misclassification error*:
$$Error(t) = 1 - \max_i [p(i|t)]$$

c is number of classes. $p(i|t)$ is the fraction of records belonging to class i at a given node t.

The main idea of the best split selection is as below:

1) Compute impurity measure (P) before splitting
2) Compute impurity measure (M) after splitting
   - Compute impurity measure of each child node
   - Compute the average impurity of the children (M)
3) Choose the attribute test condition that produces the highest gain:

$$Gain = P – M$$

or equivalently, lowest impurity measure after splitting (M).

In this study, 1000 tweets were manually labeled as either negative to flu shot or none negative to flu shot. We take 70% of the dataset for training and the rest for test. The classifier achieved performance of *precision* at *0.54* , *recall* at *0.32*, *f-measure* at *0.40* and *accuracy* at *0.80.*

B. Data Collection

The data in this paper was collected over all states of the United State in 2014 via the Twitter's streaming API. The tweets that are not in English were ignored as we expect to understand the tweets to ensure the algorithm correctness. Tweet messages are included with geographical location, the tweet timestamp and user's profile. We targeted to analyze each group's individual tweets to gain an insight into the reasons why people are against the flu shots. Here, we focus on textual pre-processing approach, hoping to leverage some useful features stand out from the noisy tweets.

C. Experimental Setup

To evaluate the effectiveness of our proposed method, we tried to reason *"why people are reluctant to take flu shots"?*. Using the supervised CART classifier, we separated tweets into two categories: *negative flu shot* and *none-negative flu shot.*

- *Negative flu shot*: Contain tweets that are reluctant to take flu shots.
- *None-negative flu shot*: Contain tweets that are either supportive or neutral to flu shots.

After the two categories (topics) were identified, we used opinion word frequency and word cloud visualization to highlight reasons for each topics. The high frequency of the opinion words could reveal the main reasons that people are reluctant to flu shots. In addition, we used word co-occurrence and correlation to ensure the opinions are actually connected with flu shots. Last, to make sure the opinions stand for different individuals, we removed all duplicated tweets, re-tweets and only one tweet message is taken into account for each Twitter's user. To evaluate the effectiveness of our normalization approach, we compared the solution with the results when not incorporating the normalization steps.

D. Implementation

The proposed approach was implemented in Python 3.6. Stanford natural language processing library [26] was used to parse sentences and performed features and opinions extraction. After tweet messages were pre-processed, the evaluation was done in R with the support of tm (text mining) package. The preporcessing tool is open source and can be accessible at *"https://github.com/litpuvn/flu-shot"*, which designed to serve as text pre-processing package in any textual analysis project.

E. Experimental Results

To find reasons why people hold negative attitude to flu shots, we compared word frequency and word correlation on our dataset with and without the proposed processing model. As we know, the word frequency can be used to measure the sentiment strength depending on its value. On the other hand, word correlation can explain the opinion towards a certain subject in the semantic context. By analyzing the two aspects, it will allow us to better understand whether the subject of the adversion is related to flu shot and how strong the word pairs appeared in the target domain.

1) Word frequency comparison: We visualized word frequency as horizontal bar chart with sorted frequency as descending order. Verbs and adjectives were treated as opinions. Only opinion words that have appeared at least five times in the tweets are displayed in the visualization. Figure 6(a) and Figure 6(b) present word frequency for traditional preprocessing model and the proposed method, respectively. In Figure 6(a), we can see that people mentioned more about *sick, hurts, paralyzed* that sound reasonable for being reluctant to flu shot. It is worth to note that there is a term *"effective"* which is supposed to supportive, however, it appeared in the negative tweets dataset. We also see on the Figure repeated words with the same meaning: *hurts* and *hurt*. This will demote the importance of the feature hurt as one of the top reasons to be against the flu shot. In summary, the traditional model introduces confusion when it shows conflicting reasons and not being able to highlight the most influencing reason when avoiding flu shots, such as *hurt*. On the other hand, the proposed model strengthens the feature selection and promotes the term *"hurt"* into a dominant reason (17 times occurred) for negative flu shots tweets. In addition, the reason *effective* in traditional model now becomes *ineffective* and thus prevents confusion in understanding the cause. Moreover, we found a new feature *die* in the early model, whose appears less than five times, now in the bar chart. In other words, the hidden reason *die* is now visible and accounts for an important factor that causes people scared of flu shots.

2) Word co-occurrence and correlation comparison: To ensure the above opinions are true reasons related to flu shots, we visualized co-occurrences of two connected terms in every tweet for comparison. Word pairs that appeared

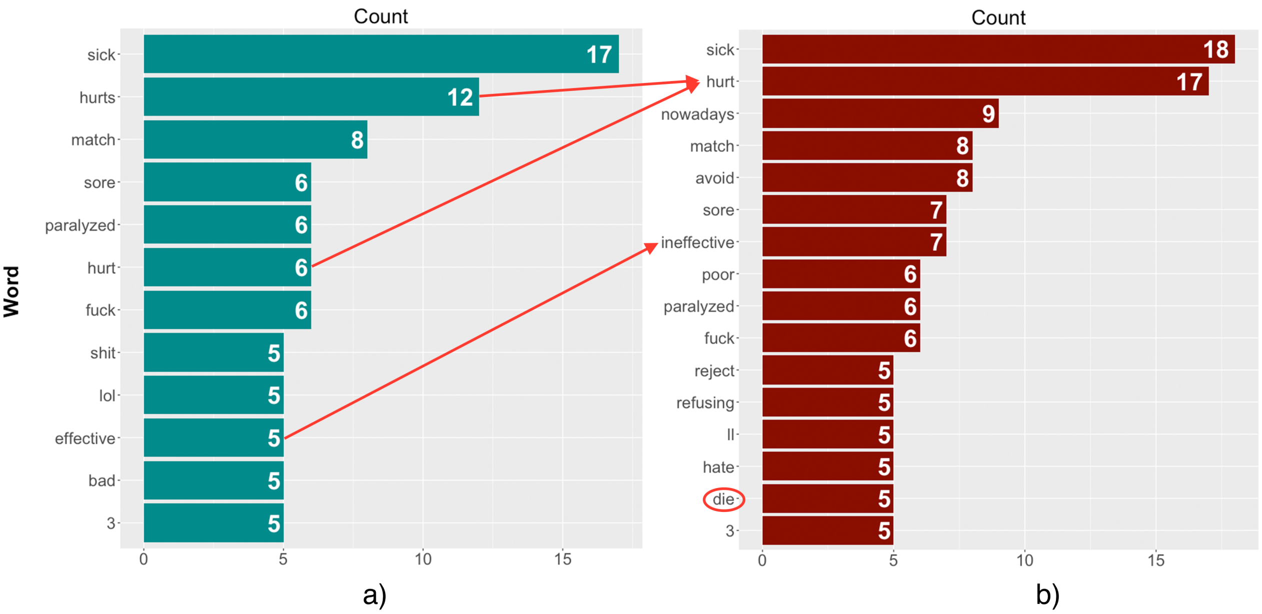


Figure 6. Word frequency used in (a) No pre-processing; (b) proposed pre-processing model

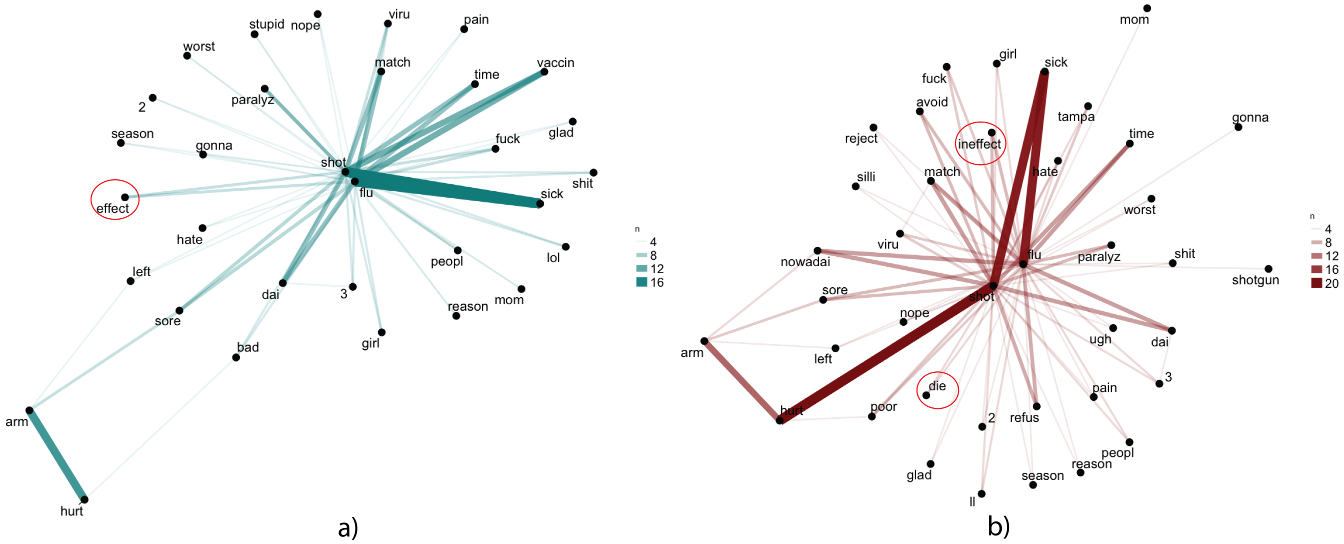


Figure 7. Word correlation used in (a) no preprocessing; (b) proposed preprocessing model

more than four times were displayed. Figure 7(a) and figure 7(b) present word correlation for traditional model and the proposed model, respectively. We can see that *flu* and *shot* are closely correlated to other terms so that they are in the middle of the representation. The stronger the relations, the thicker their connected lines. We can quickly verify that flu shots cause *hurt* in people's *arm* and *sick* in both figures with thickest connected lines. Besides, the proposed model shows that flu shots have correlation with *die*, and tell us *die* is one of actual reasons against flu shots.

V. Discussion

In this paper, we proposed a natural language processing approach in the pre-processing phase of textual analysis to enhance the reasoning output. We also demonstrated its effectiveness by answering why people are reluctant to get flu shots. The proposed method has shown the importance of text preprocessing, especially for noisy datasets like Twitter data. In addition, the word cloud visualization of the processed data also exposes stronger reasons against flu shots by highlighting important terms. There are no similar words that appears in the visualization. This has confirmed again that preprocessing is not only to reduce the amount of data but also contribute to the accuracy of the reasoning models. For future work, we will pack the language normalization tool to make it more distributable for both R and Python packages as our objective is to enhance text analysis.

References

[1] F. Jin, R. P. Khandpur, N. Self, E. Dougherty, S. Guo, F. Chen, B. A. Prakash, and N. Ramakrishnan, "Modeling mass protest adoption in social network communities using geometric brownian motion," in *Proc. KDD'14*. ACM, 2014, pp. 1660–1669.

[2] F. Jin, F. Chen, R. P. Khandpur, C.-T. Lu, and N. Ramakrishnan, "Absenteeism detection in social media," in *Proc. SDM'17*. SIAM, 2017, pp. 606–614.

[3] F. Jin, W. Wang, L. Zhao, E. Dougherty, Y. Cao, C.-T. Lu, and N. Ramakrishnan, "Misinformation propagation in the age of twitter," *Computer*, vol. 47, no. 12, pp. 90–94, 2014.

[4] F. Jin, "Algorithms for modeling mass movements and their adoption in social networks," Ph.D. dissertation, Virginia Tech, 2016.

[5] J. Hu and B. Zhang, "Product recommendation system," *CS224W Project Report*, 2012.

[6] F. Jin, W. Wang, P. Chakraborty, N. Self, F. Chen, and N. Ramakrishnan, "Tracking multiple social media for stock market event prediction," in *Industrial Conference on Data Mining*. Springer, 2017, pp. 16–30.

[7] F. Jin, N. Self, P. Saraf, P. Butler, W. Wang, and N. Ramakrishnan, "Forex-foreteller: Currency trend modeling using news articles," in *Proc. KDD'13*. ACM, 2013, pp. 1470–1473.

[8] D. M. Christopher, R. Prabhakar, and S. Hinrich, "Introduction to information retrieval," *An Introduction To Information Retrieval*, vol. 151, p. 177, 2008.

[9] H. Topi and A. Tucker, *Computing handbook: Information systems and information technology*. CRC Press, 2014, vol. 2.

[10] Z. Hai, K. Chang, and J.-j. Kim, "Implicit feature identification via co-occurrence association rule mining," *Computational Linguistics and Intelligent Text Processing*, pp. 393–404, 2011.

[11] K. Weinberger, A. Dasgupta, J. Langford, A. Smola, and J. Attenberg, "Feature hashing for large scale multitask learning," in *Proc. ICML'09*. ACM, 2009, pp. 1113–1120.

[12] T. Nasukawa and J. Yi, "Sentiment analysis: Capturing favorability using natural language processing," in *Proceedings of the 2nd international conference on Knowledge capture*. ACM, 2003, pp. 70–77.

[13] M. Collins and N. Duffy, "Convolution kernels for natural language," in *Advances in neural information processing systems*, 2002, pp. 625–632.

[14] M. Johnson, "How the statistical revolution changes (computational) linguistics," in *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?* Association for Computational Linguistics, 2009, pp. 3–11.

[15] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proc. ICML'08*. ACM, 2008, pp. 160–167.

[16] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *Journal of Machine Learning Research*, vol. 12, no. Aug, pp. 2493–2537, 2011.

[17] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," in *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*. Association for Computational Linguistics, 2002, pp. 79–86.

[18] P. Domingos and M. Pazzani, "On the optimality of the simple bayesian classifier under zero-one loss," *Machine learning*, vol. 29, no. 2, pp. 103–130, 1997.

[19] A. L. Berger, V. J. D. Pietra, and S. A. D. Pietra, "A maximum entropy approach to natural language processing," *Computational linguistics*, vol. 22, no. 1, pp. 39–71, 1996.

[20] K. Nigam, J. Lafferty, and A. McCallum, "Using maximum entropy for text classification," in *IJCAI-99 workshop on machine learning for information filtering*, vol. 1, 1999, pp. 61–67.

[21] J. Read, "Using emoticons to reduce dependency in machine learning techniques for sentiment classification," in *Proceedings of the ACL student research workshop*. Association for Computational Linguistics, 2005, pp. 43–48.

[22] E. Kouloumpis, T. Wilson, and J. D. Moore, "Twitter sentiment analysis: The good the bad and the omg!" *Icwsm*, vol. 11, no. 538-541, p. 164, 2011.

[23] A. Abbasi, H.-c. Chen, and A. Salem, "Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums," *ACM Trans. Inf. Syst.*, 2008.

[24] A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining." in *LREc*, vol. 10, no. 2010, 2010.

[25] "Pyenchant - a spellchecking library for python." [Online]. Available: http://pythonhosted.org/pyenchant/

[26] D. Klein and C. D. Manning, "Accurate unlexicalized parsing," *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, pp. 423–430, 2003.

[27] S. Schuster and C. D. Manning, "Enhanced english universal dependencies: An improved representation for natural language understanding tasks," *LREC*, 2016.

[28] "Decision tree." [Online]. Available: https://en.wikipedia.org/wiki/Decision$_tree_learning$