Sentiment Analysis of Soldiers' Tweets - Comparison with civilians (TBC)

Sumit Mukhija, Rachit Rastogi,
Chao Chen, Chen Wang, Chetan Prasad
School of Computer Science and Statistics, Trinity College Dublin
{mukhijas, rrastogi, chenc1, wangc5, cprasad}@tcd.ie

April 14, 2020

Abstract

The concern to veterans' mental health should be made. Existing works show that mental health changes caused by wars can be reflected in linguistic features of the social media texts. In order to detect and compare those changes we collected data from 20 soldiers' tweets and examined them with a list of positive and negative adjectives to identify the polarity and do a comparison with normal users' tweets. The total counts of tweets vary from 57 to 39,000. We identified the difference between normal users and soldiers and we did a close look to the result with discussion.
Keywords: Twitter, tweet, sentiment, emotion, soldier, SentiWordNet, EmoLex, lexicon

1 Introduction

Social media platforms and microblogging websites are some of the most popular online stages for people to express their views. Twitter, undeniably is one of the leading applications in this assortment. People use Twitter to post their real-time opinions in the form of tweets. These tweets can be analyzed and certain inferences can be extracted. These inferences can subsequently be used for academic and business purposes.

One of the primary reasons that make Twitter a feasible choice is the diverse nature of the users. In this research, we intend to analyze and compare the tweets of the war-veterans and the general public. We believe wars have an impact on soldiers' psychological and emotional states. We try to prove this hypothesis by comparing their tweets to the tweets posted by the civilians.

We collect public data using Twitter API and then process and count the words with a list of positive and negative adjectives to predict the polarity of the tweets. Then we examine a randomly collected dataset to compare the difference between tweets by veterans/soldiers and civilians.

(TBC due to the experiment implementation)

The remainder of the paper is organized as follows. We examine on the literature related to the topic, with papers related to previous works on the mental health of veterans, available databases on sentiment analysis and previous works done on sentiment analysis on social media in Section 2. In Section 3, we introduce our dataset and the experiment done on the dataset, with the results we have. In Section 4, we have a deep look into the result and bring the discussion. In Section 5 and 6, we conclude and bring up future works needed for the topic.

2 Background

2.1 Previous Work on Mental Health of Soldiers

In order to make medical diagnosis for patients, psychologists often use the linguistic content and expression of patients to judge their emotional changes and mental state according to previous research in psychology and linguistic. The clinical diagnosis efficiency has been greatly improved because of the progress of science and technology, especially in computational linguistics. In addition, the wide spread of social media such as Facebook, Twitter and Instagram, has provided mental researchers with a large scale of data. Therefore, they could easily use the collected dataset and machine learning techniques for sentiment analysis.

Linguistic contents which users posted on social media have been proved to be the basis for evaluating a person's mental state (Weerasinghe et al., 2019) (Guntuku et al., 2017). However, the majority of research targets are civilians. In this paper, veterans and civilians will be regarded as research targets. Westgate in (Leonard Westgate et al., 2015) has come up with a method about evaluation of Veterans' Suicide Risk. However, this paper will concentrate on analyzing the impact of the war on veterans' mental state through comparing the tweets posted by soldiers with the twitters released by civilians. In addition, a comprehensive sentiment analysis of veterans will be summarized.

2.2 Sentiment and Emotion Analysis on Social Media

Sentiment analysis has been applied to various fields, such as the research about consumer behavior on product marketing and the analysis for political voting. The advances in Natural Language Processing and linguistic research have led to the development of different methods of sentiment analysis.

Nowadays, people tend to use social media such as Twitter to post tweets and express their opinions and emotions. Most of the tweets generated from Twitter accounts are public by default and easy to obtain online. Also, the tweets are short(limited to 280 characters) and often appears with spelling mistakes and slangs. A tweet often comes with other features like spreading tweets(retweet) from other accounts. These mentioned above make the tweets become a good channel to explore sentiment analysis.

Generally speaking, the common approaches for sentiment analysis consist two parts, including the machine learning techniques based and the lexicon based. Analyzing users' social activities and calculating linguistic features of user-generated texts are the core for the machine learning algorithm. Compared with machine learning based method, the lexicon based method is more direct and straightforward. In sentiment analysis, the typical task is finding the polarities of the given texts. The tests are probably positive, negative or neutral. Lexicon based method could recognize and analyze the words about sentiment and other emoticons and hashtags which are associated with sentiment.

Therefore, the sentiment lexicons are adopted for matching the words from tweets, thus analyzing and determining the polarities of the corpus.

Azizan et al. (2019) performed sentiment analysis on Twitter data about movie review tweets using R and lexicon-based method.They found that lexicon-based method is more effective than machine learning based method under the same calculation cost. Ray and Chakrabarti (2017) used a dictionary based method and analyzed the results at aspect level and document level to predict public's sentiment using tweets about product review.

SentiWordNet and SenticNet, as open lexicons resources, have been developed in recent years. Sentiwordnet is a lexical resource which scores a text on three premises object, positivity and negativity. It is an open-source software which is free to use and helps in extracting the sentiment of the text. Due to the high accuracy, the SentiWordNet 3.0 (Baccianella et al., 2010) will be used as lexicons in this paper.

Montejo-Ráez et al. (2012) has defined a work that uses SentiWordNet on Twitter data to identify the polarity of sentiment of the users. They extract weighted vector and use it in the SentiWordNet to determine the polarity making it an unsupervisedsolution. We will use SentiWordNet on tweets in order to find the differences between the tweets of a soldier and that of a normal user.

3 Experiments and Results

3.1 Data Collection

We use TWINT by Poldi 2017 as our data collection tool. All the data can be accessed publicly so there are no ethical considerations.

Numerous strategies considered in an attempt to procure data that belonged to the war veterans. Transcripts of podcasts and YouTube videos involving accounts of wars from the veterans, books that were written by the ex-servicemen, the public dataset that had diaries, and letters of first world war soldiers were a few sources. However, as an inference, all of these sources were highly specific to the negative aspects and impacts of war and eventually would add bias to the data.

On account of being a platform that is widely used by a large number of service-men and the civilians, Twitter was selected as the platform to extract data. Several verified Twitter pages linked to the US Army were manually analysed and a few profiles of the veterans were used as an initial set. But, we finally used the data from a verified page with the name IAVA. This is said to be the most significant association speaking to the new age of vets specifically from The United States. It was ensured that veterans belonging to all the genders are selected. There was also no division on the number of Tweets. The final set had Twitter profiles with as few as 57 tweets and as many as 65,000 tweets. The succeeding veterans were picked by scouring through the followers of the veterans in the original set based on some keywords like the army, us-army, military in their bio.

A collection from 208 veterans profile was performed which was used for our final set of experiments. It helped us to have a dataset set of 6,61,342 tweets that was used in our analysis.

Similarly, for collection the data of civilians, we opted to chooses pages such as Netflix, USA official page, The US open and more. Here also the same guidelines where followed as in the case of finding the profiles of vets, but the only difference was that, here we where selection profiles based on keywords not similar to army, us-army, military in their bio. A final list of 280 civilians usernames was selected with variations in age, sex, and the number of tweets. This ultimately added 6,00,173 tweets by the general civilians and was used in our processing.

Once we had all the data that is from both the veterans and the civilians, we merged all the data to two csv file so that is easier to process the data that we have extracted. This merged data for both where further used in our pre-processing.

3.2 Data Process and Analysis

3.2.1 Lexical & Non-Lexical Feature Summarization

Once the data is gathered and saved in the CSV format, we then start pre-processing the gathered data after which we do our analysis where we come up with a set of results to prove our hypothesis. Sentiment analysis can be broadly categorized into two kinds, based on the type of output the analysis generates. Under our processing, we are trying to label text to be under "curses" – or bad words. Take into account the tweets of all the selected veterans from our data and then run an analysis to gather information from there. Not only this, information such as the total counts of retweets, replies, likes, emojis, URLs, mentions and total words are calculated. Special elements in tweets can be referred in Table.1. This is one of the analysis that we are dong from the collected data.

This identical analysis is then run on tweets by ordinary people or civilians data as well. Finally, we will categorise the difference in the count of "curse" words to check the two sets of results which will ultimately help us identify an individual being in the state of depression.

The initial step we took under pre-processing our data were:

- Only selecting tweets that where in English Language.
- Applying rule to only select textual data.

It was then followed by removal of: (1) Retweets numbers; (2) Tokenizing.

Furthermore, the emojis where given a textual form so that we can get information from that part as well. This is because, in this everchanging world the use of emojis has increases. This converted emojis was considered as text and was used in the count of curses.

The step to remove punctuation followed by Tokenizing which is the way toward separating a goliath string into a rundown of words. NLTK (Bird et al., 2009), a python library is used for this process.

Stopwords, where pull-out, as they do not change any meaning of a sentence hence, can be ignored. Tuples were generated with each word and part of speech. Finally, the counts were extracted from the tweets of both the veterans and the civilians. And was then was compared with each other.

A abstract of the process is illustrated in Figure.1.

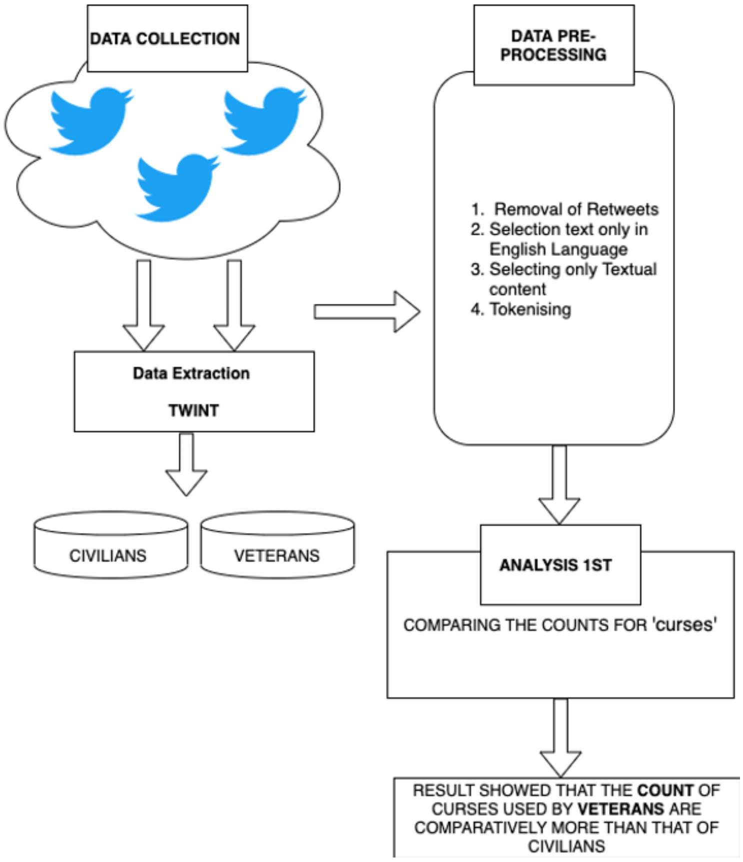

Figure 1: Process of lexical & non-lexical feature summarization

Table 1: Elements to handle when prepocessing tweets

Element	Examples	Element	Examples
URLs	http://foo.bar	Blank spaces	
Mentions to other users	@Bot	Single letter words	a b c
Hashtags	#botRise	Numbers	1994 233
Twitter reserved words	RT via	Stopwords	it I as

Hence, we compared the count of the “curses” to come up with a result that displayed whether the veterans were targeted to depression and suicidal. Here is the data that was collected from the the tweets:

- soldier stats: 'retweets': 869272, 'replies': 333204, 'likes': 2735134, 'emojis': 159583, 'urls': 250026, 'mentions': 226533, 'words': 51528493, 'curses': 28249, 'tweets': 661342
- civilian stats: 'retweets': 24102124.0, 'replies': 5919866.0, 'likes': 123551386, 'emojis': 69284, 'urls': 248854, 'mentions': 325691, 'words': 44355619, 'curses': 8983, 'tweets': 600173

The entire idea for our research was to check the mental condition of veterans or any army person to that of to a normal civilian and therefore draw a sentimental analysis. Our research done on nearly 488 people that include the civilian's as well as the vets shows that there is definitely effect of wars that can be reflected from the type of words they use while tweeting. One of the analysis shows that the number of CURSES used in by the vets are much more greater that that used by an ordinary civilians. The word count for curse word by vets accounts to be 28,249 which is much greater than that curse world count used by civilians which is at 8,983.

3.2.2 Sentiment & Emotion Analysis

Tweets are filtered and only tweets with texts originate from uses themselve remain, which means the likes and directly retweets are filtered.

The corpora are then preprocessed to remove elements mentioned in Table.1.

To notice that, when we remove numbers we try to remain the years (from 1900 to 2100), and we try not to remove punctuations and stopwords because we need to do Part-of-Speech (POS) tagging after tokenizing. Both tokenizing and POS tagging is done by NLTK (Bird et al., 2009).

We use lexicons to score the words in our corpora. SentiWordNet is used for sentiment polarity analysis and NRC Word-Emotion Association Lexicon (EmoLex) (Mohammad and Turney, 2013) based on the model of Plutchik's wheel of emotions (Plutchik, 2003) (with additional Positiveness and Negativeness) is for emotion analysis.

Once the POS tags of words are generated. We search the synonyms of words in SentiWordNet to determine the scoring for positiveness, negativeness and objectiveness by caculating means among synonyms. Meanwhile EmoLex is used to perform emotion analysis on 10 emotions. Scores of one tweet are generated caculating the means of the scores of all the words after preprocessing.

The result data applying SentiWordNet are shown in Table.2. Result produced by using EmoLex are shown in Table.3.

We also counted adjectives with top 100 frequencies in soldiers and civilians corpora, for we think that adjectives have more subjective meanings than verbs, nouns, etc. We discovered some words with more "political" meanings appear to be different in the lists of two corpora. The list of adjectives are shown in Table.4.

Table 2: Results of sentiment analysis using SentiWordNet

		Valid Cnt.	**Valid Len.**	**Possitive.**	**Negative**	**Objective.**
Soldiers	Mean	3179.54*	16.450	257.58×10^{-4}	196.10×10^{-4}	3371.7×10^{-4}
n=208	Std.	5041.70	6.6427	78.586×10^{-4}	64.386×10^{-4}	750.49×10^{-4}
Civilians	Mean	2143.66*	14.293	262.65×10^{-4}	177.39×10^{-4}	3530.5×10^{-4}
n=280	Std.	5286.12	5.2067	87.432×10^{-4}	64.786×10^{-4}	720.09×10^{-4}

4 Discussion

From Table.2 we cannot get much inference on the sentiment part, instead we find that soldiers are more likely to send long texts (see the numbers with *).

We can see from Table.3 that corpus of soldiers' tweets has more "negative" emotions like Disgust, Fear, Anger and Sadness. The corpus of soldiers' tweets is judged as negative on the whole. While civilians' corpus tend to be more possitive, with better metrics on Surprise, Anticipation, and Joy. One interesting item is Trust, from Plutchik's wheel of emotions **??** we can infer that Trust is a kind of emotion related to submission, acceptance and admiration, which is related to soldiers' loyalty obeying the commands. While Surprise is related to disaproval and distraction, which can somewhat indicating the quality of disorder among internet users.

With the list of adjectives (Table.4) we can see that soldiers are more involved in political topics.

Table 3: Results of emotion analysis using EmoLex

Soldiers: n=208					
	Trust+	**Anger+**	**Surprise**	**Joy**	**Positive.**
Mean $\times 10^{-4}$	422.84	167.17	149.99	312.51	637.50
Std. $\times 10^{-4}$	154.94	83.143	62.113	151.64	213.55
	Disgust+	**Fear+**	**Anticipat.**	**Sadness+**	**Negative.+**
Mean$\times 10^{-4}$	122.09	193.43	295.57	149.34	339.61
Std.$\times 10^{-4}$	74.795	89.380	112.40	66.053	148.79
Civilians: n=280					
	Trust	**Anger**	**Surprise+**	**Joy+**	**Positive.+**
Mean $\times 10^{-4}$	399.72	132.03	163.72	349.44	650.63
Std.$\times 10^{-4}$	170.45	80.189	108.13	224.58	269.03
	Disgust	**Fear**	**Anticipat.+**	**Sadness**	**Negative.**
Mean$\times 10^{-4}$	98.934	163.78	330.09	131.82	283.00
Std.$\times 10^{-4}$	82.884	108.23	152.86	87.180	160.08

Table 4: List of “political” adjectives with rankings and frequencies

Word	**Soldiers**	**Civilians**	**Word**	**Soldiers**	**Civilians**
military	4.46 (17th)	-	dead	1.20 (77th)	-
american	3.85 (24th)	1.20 (79th)	human	1.17 (80th)	1.01 (91st)
political	2.12 (40th)	-	local	1.16 (81st)	1.26 (72nd)
medical	1.75 (47th)	-	democratic	1.15 (83rd)	-
public	1.56 (51st)	1.32 (68th)	illegal	1.14* (85th)	-
social	1.44 (60th)	1.78 (51st)	foreign	1.14* (85th)	-
sick	1.37 (64th)	-	poor	1.10 (90th)	-
personal	1.31 (71st)	1.21 (78th)	republican	1.07 (93rd)	-

5 Conclusion

6 Future Works

Thought we did try to answer our research question by applying two different analysis, there is always scope of improvement. One area where we like to work further is to extract data depending upon the timestamp. We wanted to get extract data depending on times when the veterans returned from a war and then compare it with his own tweets that they did before the war, but this can be done in future. And definitely is an area to work forward.

References

Azizan, A., N. N. S. A. Jamal, M. N. Abdullah, M. Mohamad, and N. Khairudin (2019). Lexicon-based sentiment analysis for movie review tweets. In *2019 1st International Conference on Artificial Intelligence and Data Sciences (AiDAS)*, pp. 132–136. IEEE.

Baccianella, S., A. Esuli, and F. Sebastiani (2010). Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In *Lrec*, Volume 10, pp. 2200–2204.

Bird, S., E. Klein, and E. Loper (2009). Natural language processing with python.

Guntuku, S. C., D. B. Yaden, M. L. Kern, L. H. Ungar, and J. C. Eichstaedt (2017). Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences 18*, 43–49. Big data in the behavioural sciences.

Leonard Westgate, C., B. Shiner, P. Thompson, and B. V. Watts (2015). Evaluation of veterans' suicide risk with the use of linguistic detection methods. *Psychiatric Services 66*(10), 1051–1056. PMID: 26073409.

Mohammad, S. M. and P. D. Turney (2013). Crowdsourcing a word-emotion association lexicon. *29*(3), 436–465.

Montejo-Ráez, A., E. Martínez-Cámara, M. T. Martin-Valdivia, and L. A. U. López (2012). Random walk weighting over sentiwordnet for sentiment polarity detection on twitter. In *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*, pp. 3–10.

Plutchik, R. (2003). *Emotions and life: Perspectives from psychology, biology, and evolution.* American Psychological Association.

Ray, P. and A. Chakrabarti (2017). Twitter sentiment analysis for product review using lexicon method. In *2017 International Conference on Data Management, Analytics and Innovation (ICDMAI)*, pp. 211–216.

Weerasinghe, J., K. Morales, and R. Greenstadt (2019). "Because... I was told... so much": Linguistic indicators of mental health status on twitter. *Proceedings on Privacy Enhancing Technologies 2019*(4), 152–171.