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 Veterans

In order to make a 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 widespread 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 normal people.

In this paper, veterans will be regarded as research targets. Westgate in (Leonard Westgate et al., 2015) has come up with a method about the evaluation of veterans' suicide risk. This paper will concentrate on analyzing the impact of the war on veterans' mental state through the Twitters posted by themselves before and after the war instead of focusing on the prediction of the suicide risk of veterans. In addition, the comparison with the twitters released by ordinary users will be presented. Finally, comprehensive sentiment analysis of veterans will be summarized.

2.2 Sentiment and Emotion Analysis on Social Media

Emotion and sentiment are treated as different concepts in psychology. The definition of "emotion" is a complex psychological state, which plays an important role in operating motivators. For "sentiment", it is created based on emotion to refer to a mental attitude. A survey providing more information can be referred to by (Yue et al., 2018).

In sentiment analysis, the typical task is finding the polarities of the given texts. The tests are probably positive, negative or neutral. The approach is often counting the word using and produce scores due to the lexicons.

There are commonly two approaches - analysing users' social activities and calculate linguistic features of user-generated texts. The sentiment analysis mainly focuses on short texts(tweets) generated from Twitter accounts, since most of the data is 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 methods mentioned above make the analysis on tweets a paradigm to explore.

SentiWordNet makes use of Opinion Mining, which is understanding the opinion of text more than the topic (Esuli and Sebastiani, 2006). Sentiwordnet is a lexical resource which scores a text on three premises object, positivity and negativity. Synsets are the building blocks of the Sentiwordnet, they form a wordnet and the wordnet is associated with the three scores. The three scores help determine how objective, positive and negative the text is. Sentiwordnet is an open-source software which is free to use and helps in extracting the sentiment of the text. The latest version of this is SentiWordNet 3.0 (Baccianella et al., 2010) which is being used in this project. The latest version of Sentiwordnet uses an updated Wordnet compared to the older version. The algorithm used is updated to include random walk step to refine the scores. There is also considerable improvement in the accuracy of Sentiwordnet 3.0. It is used in numerous projects for the analysis of reviews and other related matters to understand whether the text is subjective or objective.

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 are going to be using SentiWordNet on tweets in order to understand the differences between the tweets of a soldier and that of a normal user.

3 Experiment and Results

3.1 Experiment Setup

3.1.1 Data Collection

3.1.2 TBC

3.1.3 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:

Table 1: Elements to be removed when proprocessing

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		

When we remove numbers we try to remain the years (from 1900 to 2100). 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 corpus. SentiWordNet is used for sentiment polarity analysis and NRC Word-Emotion Association Lexicon (EmoLex) (Mohammad and Turney, 2013) 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, and 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}

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 word rankings and frequencies

Word	**Soldiers**	**Civilians**	**Word**	**Soldiers**	**Civilians**
military	4861 (17th)	-	dead	1302 (77th)	-
american	4193 (24th)	1101 (79th)	human	1273 (80th)	925 (91st)
political	2305 (40th)	-	local	1260 (81st)	1152(72nd)
medical	1914 (47th)	-	democratic	1248 (83rd)	-
public	1704 (51st)	1209 (68th)	illegal	1246* (85th)	-
social	1566 (60th)	1628 (51st)	foreign	1246* (86th)	-
sick	1488 (64th)	-	poor	1202 (90th)	-
personal	1428 (71st)	1111 (78th)	republican	1162 (93rd)	-

4 Discussion

5 Conclusion

6 Future Works

References

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.

Esuli, A. and F. Sebastiani (2006). Sentiwordnet: A publicly available lexical resource for opinion mining. In *LREC*, Volume 6, pp. 417–422. Citeseer.

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.

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.

Yue, L., W. Chen, X. Li, W. Zuo, and M. Yin (2018). A survey of sentiment analysis in social media. *Knowledge and Information Systems*, 1–47.