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 13, 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:

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 Available Database on Sentiment Analysis

Sentiment analysis is a very useful method for analysing the sentiment of an article, tweets or reviews. The advances in Natural Language processing and linguistic research have led to the development of different methods of sentiment analysis. The sentiment analysis is an integral part of our work and it is the fulcrum on which our hypothesis hangs on. Determining not just the sentiment of a text but even the topic (Lin and He, 2009) on which it is written on is one of the interesting works in this field. Another work makes use of the lexicons in the text and word dictionaries to extract the sentiment behind it (Maas et al., 2011). There is another work that trains a model to learn not just the words but to also capture the sentiment behind it (Taboada et al., 2011).

In this work, we do sentiment analysis on tweets and we compare the tweets of two types of users. We utilise a simple model which basically just classifies the tweets into Happy and sad/depressing sentiments. This is sufficient considering our work is to identify the difference in the mental states between the soldiers who have served in wars and common people.

2.3 Sentiment 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.

The combination of machine learning and data mining techniques are the key part of sentiment analysis on social media. 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 mentioned above makes the analysis on tweets a paradigm to explore.

Barbosa and Feng (2010) produced an approach to automatically analyse sentiment on tweets with metainformation as retweets, hashtags, replies, punctuations and emoticons. They also use sources of noisy labels in their training data to test the robustness of the model.

Agarwal et al. (2011) performed sentiment analysis on Twitter data using POS- specific prior polarity features, the lexicon features and also the meta features. They use a tree structure to represent tweets and used a partial tree kernel to measure the similarity between two set of tweets.

Liang and Dai (2013) developed a Unigram Naive Bayes model for sentiment analysis on Twitter data. The χ^2 feature extraction method and Mutual Information were used to delete unwanted features, and then predictions were made on the tweets whether they were positive or negative. They found that the combination of prior polarity of words and their POS tags affects the most.

Bhavsar and Manglani (2019) used a dataset from Kaggle and classify the people emotions based on positive and negative reviews. The model they produced can perform well on a large dataset.

The International Workshop on Semantic Evaluation(also known as SemEval) started holding tasks related to sentiment analysis on social media since 2013. The rankings of each tasks are made public also with solutions. A list of task can be found in Table.1.

Table 1: The tasks of SemEval on sentiment analysis of tweets

Workshop	**Task**
SemEval-2013	Task 2: Sentiment Analysis in Twitter (Nakov et al., 2013)
SemEval-2014	Task 9: Sentiment Analysis in Twitter (Rosenthal et al., 2014)
SemEval-2015	Task 10: Sentiment Analysis in Twitter (Rosenthal et al., 2015) Task 11: Sentiment Analysis of Figurative Language in Twitter (Ghosh et al., 2015)
SemEval-2016	Task 4: Sentiment Analysis in Twitter (Nakov et al., 2016) Task 6: Detecting Stance in Tweets (Mohammad et al., 2016)
SemEval-2017	Task 4: Sentiment Analysis in Twitter (Rosenthal et al., 2019) Task 6: #HashtagWars: Learning a Sense of Humor (Potash et al., 2017)
SemEval-2018	Task 1: Affect in Tweets (Mohammad et al., 2018) Task 3: Irony Detection in English Tweets (Van Hee et al., 2018)
SemEval-2019	Task 5: HatEval: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter (Basile et al., 2019) Task 6: OffensEval: Identifying and Categorizing Offensive Language in Social Media (Zampieri et al., 2019)

3 Experiment and Results

3.1 Experiment Setup

3.1.1 Data Collection

3.1.2 TBC

3.1.3 Sentiment Analysis

We use a lexicon called SentiWordNet for scoring the words in our corpus. The corpus is first preprocessed to remove elements mentioned in Tab.2:

Table 2: 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		

We try not to remove punctuations and stopwords because we need to do Part-of-Speech(POS) tagging after tokenizing. When we remove numbers when try to remain the years (from 1900 to 2100). Once the POS tags of words are generated. We look up the words

3.1.4 Emotion Analysis

3.2 Results

4 Discussion

5 Conclusion

6 Future Works

References

Agarwal, A., B. Xie, I. Vovsha, O. Rambow, and R. J. Passonneau (2011). Sentiment analysis of twitter data. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, pp. 30–38.

Barbosa, L. and J. Feng (2010). Robust sentiment detection on twitter from biased and noisy data. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING' 10, USA, pp. 36–44. Association for Computational Linguistics.

Basile, V., C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. R. Pardo, P. Rosso, and M. Sanguinetti (2019). Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pp. 54–63.

Bhavsar, H. and R. Manglani (2019). Sentiment analysis of twitter data using python. *International Research Journal of Engineering and Technology (IRJET) Mar 2019e-ISSN*, 510–511.

Ghosh, A., G. Li, T. Veale, P. Rosso, E. Shutova, J. Barnden, and A. Reyes (2015). Semeval-2015 task 11: Sentiment analysis of figurative language in twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pp. 470–478.

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.

Liang, P. and B. Dai (2013, 6). Opinion mining on social media data. In *2013 IEEE 14th International Conference on Mobile Data Management*, Volume 2, pp. 91–96.

Lin, C. and Y. He (2009). Joint sentiment/topic model for sentiment analysis. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM' 09, New York, NY, USA, pp. 375–384. Association for Computing Machinery.

Maas, A. L., R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts (2011). Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT' 11, USA, pp. 142–150. Association for Computational Linguistics.

Mohammad, S., F. Bravo-Marquez, M. Salameh, and S. Kiritchenko (2018). Semeval-2018 task 1: Affect in tweets. In *Proceedings of the 12th international workshop on semantic evaluation*, pp. 1–17.

Mohammad, S., S. Kiritchenko, P. Sobhani, X. Zhu, and C. Cherry (2016). Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pp. 31–41.

Nakov, P., A. Ritter, S. Rosenthal, F. Sebastiani, and V. Stoyanov (2013). Semeval-2013 task 2: Sentiment analysis in twitter. In *SemEval@NAACL-HLT*, USA, pp. 312–320. Association for Computational Linguistics.

Nakov, P., A. Ritter, S. Rosenthal, F. Sebastiani, and V. Stoyanov (2016). Semeval-2016 task 4: Sentiment analysis in twitter.

Potash, P., A. Romanov, and A. Rumshisky (2017). Semeval-2017 task 6:# hashtagwars: Learning a sense of humor. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pp. 49–57.

Rosenthal, S., N. Farra, and P. Nakov (2019). Semeval-2017 task 4: Sentiment analysis in twitter.

Rosenthal, S., S. M. Mohammad, P. Nakov, A. Ritter, S. Kiritchenko, and V. Stoyanov (2015). Semeval-2015 task 10: Sentiment analysis in twitter.

Rosenthal, S., A. Ritter, P. Nakov, and V. Stoyanov (2014). Semeval-2014 task 9: Sentiment analysis in twitter. *ArXiv abs/1912.02990*.

Taboada, M., J. Brooke, M. Tofiloski, K. Voll, and M. Stede (2011). Lexicon-based methods for sentiment analysis. *Computational linguistics 37*(2), 267–307.

Van Hee, C., E. Lefever, and V. Hoste (2018). Semeval-2018 task 3: Irony detection in english tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pp. 39–50.

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.

Zampieri, M., S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar (2019). Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). *arXiv preprint arXiv:1903.08983*.