Happy or sad emotion in online user content

Huub Exel, RUG Information Science Student

Abstract— **In the Netherlands there is a saying which says that the people in the North are less happy and friendly than people in the south. In this paper is going to research whether this might be true or this might be false. A python code is developed to check whether certain good or bad words in tweets are present and according to that the tweets are being graded in a system either good, neutral or bad. The results are being analysed and discussed according to the chi square test results.**

Keywords — **Emotion, Language processing, north vs south, Netherlands**

I. Introduction

IN the Netherlands there is a well-known saying that people who live in the North of the Netherlands are way more stubborn, unfriendly, cold and unhappy as well, than people who live in the south of the Netherlands. This could have many reasons. The sun could be shining more, it is more mountainous and there is more nature in the South. Non of those reasons matter in this paper, this is not an attempt to find out why people in the north may be more unhappy and unfriendly than people in the south. This paper is written to check and see if the saying is actually true or not. Is there actually a significant difference or is the saying somewhat false. It might not be that big of difference as people say that it is.

If someone would like to come and live in the north of the Netherlands, small little town with less than 10.000 are almost never considered as a good place to live, partially because people think that the original residents are not happy and friendly people. In this paper the information gap, whether people in the north are less happy and friendly than people in the South, is going to be filled more than it was before this paper.

In this paper the aim is to develop a better understanding in how happy or sad the people of the north and the south of the Netherlands really are, this by analyzing tweets from people who live in those areas. This paper will contribute by giving the code and outcome necessary to get the better understanding. The whole data-set will become available for anyone who wants to do research on it. Prior works use mostly Neutral Networks to accomplish their research however in this paper we only use the data.

In II the related work will be discussed, what other works do to identify emotion in tweets. Where the data comes from, the content of a tweet, what is included and what is excluded in this work is will be discussed in III. In IV the primary dataset, the tweets, will be discussed in further detail. V discusses the other datasets that are needed to make this work possible. The methodology is discussed in VI, how the author carried out his research. In VII the experiments themselves, the rules of the experiments and the results of the experiments are explained. And last but not least at VIII the conclusions and future work is discussed, what could be done different next time.

The question that the author is going to try to answer in this paper is: Is there a difference in happiness between people who live in the North of the Netherlands and people who live in the South of the Netherlands The hypothesis is that there is not going to be a difference between the happiness of the two groups.

II. Related Work

Many related works have been done on the topic of trying to detect emotion in tweets, most of those use Natural Language Processing (NLP) Neural Networks to measure the emotion in the tweet. In [1] authors fetched tweets from Twitter to create a dataset. After this they obtain target based extended features models. They train different supervised classifiers such as Artificial Neural Networks or a Maximum Entropy. In [2] the authors made a corpus with Twitter tweets and used corpus annotation study on that corpus. Support Vector Machines (SVM) were used as the learning model. The features they selected include things as unigrams, bigrams, pronouns, adjectives and things as Word-net Affect emotion lexicon. In [3] the authors put the data in seven classes of emtions Anger, Disgust, Fear, Guilt, Joy, Sadness, Shame. They tokenize their data and use a stemming algorithm to make their data more usable. The features used are unigrams, bigrams and trigrams. A special scheme is applied to score the n-grams. The Multinomial Naïve Bayes is used as a classifier which is trained with the highest scoring n-grams and accuracy is tested with different feature sets.

These works are more advanced than the work that will be done in this paper.

III. Data

The data that is used in this paper is extracted from Karora which is a computer at the RUG that collects all tweets that are, according to it's algorithm, Dutch. The data exist out of the tweet and the metadata that needs to be researched which can vary from the id of the user that send the tweet, to where he or she lives. Unfortunately not all of the tweets have all metadata filled in, or all metadata filled in with information useful for this research (everyone can fill in whatever he or she likes so even non places could be filled in). For this work the focus lies on people that have filled in where they live, and only those that have filled in that they live in the north (Drenthe, Groningen, Friesland) or south (Zeeland, Noord-Brabant, Limburg) of the Netherlands. Anyone who does not live in those provinces is excluded from this paper. The coordinates of where the tweets where sent where also collected by the code written however they are not used. Places data files and good, bad word data files are used but are discussed elsewhere.

Variables: The dependant variables are the positive tweets and negative tweets. The independent variables are the north- and south of the Netherlands

IV. Primary Dataset

The primary dataset consist of all Dutch tweets from January 2018 up to and including all Dutch tweets from April 2018. Their is not an indication of how many percent will be happy tweets and how many percent will be sad tweets. These tweets also have as metadata the place where the user, who send the tweet, lives and the coordinates of where the tweets are send. These two are not filled in for every tweet.

V. Places, good and bad words datasets

To see whether a tweet is from the north or south or whether its good or bad, data-sets are needed to check those properties. In the good and bad words files are only strictly good and strictly bad words. For example, the word "vernederd" (humiliated) is never used in a good way. If this was used in a good way than the person who used it, like it was a good word, made a mistake. The same goes for words that are strictly good, for example, the word "blij" (happy) cannot be used in a bad way. You could say "Ik ben niet blij" (I am not happy), how that is handled in this paper is discussed elsewhere. The words in both files have been checked by the author of this paper, whether the word is truly good or truly bad, and if that was not the case, the word would get deleted out of the file.

There is also a north places file and a south places file, which contain either all north places of the Netherlands (including the Frisian names for the Frisian places) or all south places of the Netherlands. These are made by getting a datafile with all places of the Netherlands in it and using a python file (cleaner.py) on it.

VI. Methodology

For this work the author extracted tweets from Karora which collects all of the Dutch tweets, the author did this in such a way that every day is separately extracted. After this process is done, the author put every file into a directory according to the month that the tweets were from. Lastly he put all files together in 1 big data file. After this, all places of the Netherlands were put into one file and filtered on north and south by a python file (cleaner.py). For the last bit of data, a good words file and a bad words file was made and scanned to see if the words in it were strictly good and bad.

The places files, the good- and bad words files were imported into the main python file (worker.py) and put into sets to make them more efficient to work with. The different file path names were created by using formatting and were put into 1 list to be used later on. For every file in the data file there were certain things that had to be done. First of all, the metadata of the tweet had to be checked where the tweet came from. The place metadata of the tweet was tokenized and checked with the places in the set of the north and south, if the tweet came from the north or the south of the Netherlands, the tweet was temporarily saved (north separated from south) and, if all the data from that day was checked, passed on to the next function. The tweets were than tokenized by the NLTK tokenizer in this part of the code and passed on to the next function. In this function the most important part of the process happened, checking whether the tweet is happy or sad. The tweets could be placed in 3 categories, good, bad or neutral. Two counters are introduced, one for good tweets and one for bad tweets. If the tweet is good, the good tweet counter would get plus one and vice versa for a bad tweet and that counter.

After all data in one day was fully checked the number of the two counters would be added to a total counter for either good or bad tweets. These are the final counters and this is were the calculations will be based on.

If there is anything unclear about the code `https://github.com/hacxpiont/Introduction_to_research_methods/tree/main/FinalProject`, the comments in the code itself will explain in even greater detail how certain things work.

VII. Experiments

The experiment was done according to the Methodology. Some things that do have to be considered, earlier on in this paper the system of a tweet

either being good, bad or neutral was introduced. How this works is the following, every word in the lists of either the good or bad words is worth one point. If a tweet has only 1 word of either the good or the bad words than it means that tweet is good or bad (depending on the word). If the tweet has one good and one bad word or two good words and two bad words and so on, the score of the tweet is 1-1 (or 2-2, etc.), which means that the tweet is considered neutral, it is neither good nor bad. A good example of this would be the sentence "het was leuk, maar niet bijzonder, ik zou niet nog een keer gaan." (It was fun but not exceptional, I would not go agian), both "leuk" and "niet" are worth 1 point, making the score 1-1. In the sentence, "It was fun but not exceptional, I would not go again.", the person already expected it to be somewhat fun otherwise he would not say "but not exceptional". For example when someone goes to a comedy night at the local bar it is expected to be fun, however if it is not super fun it is not exceptional making this sentence neutral. If the tweet has a 2 or more advantage over the other type it is still considered good or bad.

In this paper we do not use the most recent tweets available, this is because in the last year and in the year before that the world was stricken by Covid-19 which could make the outcome of this research uncontrollably different. The last entirely Corona free year is 2018. The author thinks that four months of data is lengthy enough for a research of this type.

How the tweets and places of the tweets are checked is by first tokenizing them and than running them over the set of places, the good words and the bad words. The author also tried to match every substring of what was filled into the place to match that with places in the set, however there are a lot of places in the north and south of the Netherlands which have only three letters as there place name so what happened was a lot of substrings ended up getting matched without it being the intention that those got matched.

The results of the experiments are shown in the tables below.

The chi squared is used here because the relationship between two categorical variables needs to be assessed in this work.

Frequencies

Emotion of Tweets			
	Good	**Bad**	**Total**
North	225378	264243	489621
South	566790	608427	1175217
Total	792168	872670	1664838

Proportions

Emotion of Tweets			
	Good	**Bad**	**Total**
North	0.1354	0.1587	0.2941
South	0.3404	0.3655	0.7059
Total	0.4758	0.5242	1

Proportions per row

Emotion of Tweets		
	Good	**Bad**
North	0.4603	0.5397
South	0.4823	0.5177

Proportions per column

Emotion of Tweets		
	Good	**Bad**
North	0.2845	0.3028
South	0.7155	0.6972

Expected frequencies

Emotion of Tweets			
	Good	**Bad**	**Prop.**
North	232977	256652	0.2941
South	559191	616018	0.7059
Total	792168	872670	

Chi square score

Emotion of Tweets			
	Good	**Bad**	**Prop.**
North	247.59	224.75	489621
South	103.15	93.64	1175217
Total	792168	872670	

Outcomes of the chi square test are the following:
The chi square score is 669.13
Degrees of freedom is 1 (df is 1)
p-value is smaller than 0.0001
Cramer's V is 0.02

The p-value is really utterly low, much lower is not possible. With an α is 0.05, the p-value is much lower. The null hypothesis says that there is not a difference between the north and the south when it

comes to good or bad words. The alternative says that there is a difference between the north and the south when it comes to good and bad words. In this case 0.0001 is lower than 0.05 that means that the alternative hypothesis can be accepted. The result is significant. There is a big side-note here, the Cramer's V is 0.02 which is Negligible. So if the outcome is useful is debatable.

VIII. Conclusions and Future work

As the Cramer's V already showed us, the result of the effect size is negligible. the author thinks that there is a tiny difference between the north and south of the Netherlands but not big enough to discuss. Checking to see if certain words are the same as in a set with for example good words is not the best way to see whether tweets are happy or sad. In future works Neural Networks could be used to obtain the results of this work. The coordinates could also be used in the future to see whether something is in the north or south. Future works could try and use the context of the tweet to see whether its a good or bad tweet. Different kinds of happiness and sadness could be added.

Acknowledgements

The author would like to thank Mr. Caselli for teaching the Introduction to Research Methods course to the author which made it possible to write this paper the way it is.

References

[1] M. Anjaria and R. M. R. Guddeti, *"Influence factor based opinion mining of Twitter data using supervised learning"*, IEEE, Sixth International Conference on Communication Systems and Networks (COMSNETS), 2014.

[2] M. Mohammad R. C. Balabantaray and N. Sharma, *"Multi-class Twitter Emotion Classification: A New Approach", International Journal of Applied Information Systems"*, vol. 4, 2012.

[3] P. Vinod B. Thomas and K. A. Dhanya, *"Multiclass Emotion Extraction from Sentences," International Journal of Scientific amp; Engineering Research"*, vol. 5, 2014.