CS7602-Machine Learning Techniques

Course Project

TITLE: Understanding Personality through Social Media

Team Members

Shri Shruthi-2015103022

Hari Krishna Nagarajan-2015103017

Abstract:

In this paper, we study the relationship between language use on Twitter and personality traits. Specifically, we want to know how various linguistic features correlate with each personality trait and to what extent can we predict personality traits from language. We gather personality data from Myers-Briggs Type Indicator (MBTI) personality test which contains thinking, feeling, sensation, intuition, introversion, extroversion, judging and perceiving. Using the 1000 users in our dataset, we collect most recent tweets from them and design three categories of feature, namely bag of n-grams, Twitter POS tags, and word vectors to explore the most related linguistic features for different personality traits. Analysis of these features provide insights of language use for different personalities. For instances, extroverts tend to use hashtag and phrases like ”so proud”, ”so excited”, and ”can’t wait”. People who like to use emoticon are more likely to be Sensing and Feeling personality type. Moreover, we investigate the predictive power of individual features and combined features in our analysis.

1. Introduction

Personality has been studied extensively in social science and psychology as it reflects the way people behave and react in online social media and in the society. Previous studies showed that personality significantly correlated with several real-world behaviors which makes it important in providing personalized services.People are increasingly using social media platforms, such as Twitter, Facebook, and Pinterest, to share their thoughts and opinions with their friends or people who are interested. Such scale of social media platforms provide us with a unprecedented opportunity to understanding psychological attributes on a large user base.Specifically, we aim to find the linguistic features that distinguish people with different personality types and explore how these features can be explain by personality. Further, using the these features we want to understand the degree to which we can predict personality traits from social media language.

However, little research has touch upon understanding personality through social media because of a few reasons.Due to the format of Twitter its very hard to collect data as it has Twitter specific-Language like #(Hastags) and 140 characters messages which makes content richer but makes it hard for analysis by typical linguistic analysis tool. Also collecting personal dat ais very hard as the subjects may need to fill out a questionnaire and this approach makes it harder for us to scale it up.

However in this paper, we try to solve the aforementioned problem by designing new richer linguistic analysis tools which can extract language feature in social media context and introducing a mechanism to automatically extract personality from text in social media.

2. Related Work

Social media sites are now the most popular destination for Internet users, providing social scientists with a great opportunity to understand online behavior. There are a growing number of research papers related to social media, a small number of which focus on Analyzing and predicting personality that has recently attracted more and more attention in research community. Golbeck et al. used tweets on Twitter to extract Linguistic Inquiry and Word Count (LIWC) features, MRC language features, Twitter use, structural, and sentiment features to predict the personality traits . Sumner et al. explored the extent to which it is possible to determine antisocial personality traits based on Twitter use. Quercia et al. studied the relationship between personality traits and five types of Twitter users: listeners, popular, highly-read, and two types of influentials. These studies, however, were mostly based a small samples of hundreds or thousands of users that could be prone to bias in the dataset.

However, this work only looked at the general language use instead of online-specific language use which ignores an important part of online behaviors. Also, this work has not investigate the relationship between word vector features and personalities which are proved to be the best predictive feature in our analysis.

4. System Design

The following diagram summarizes the steps involved in system design :

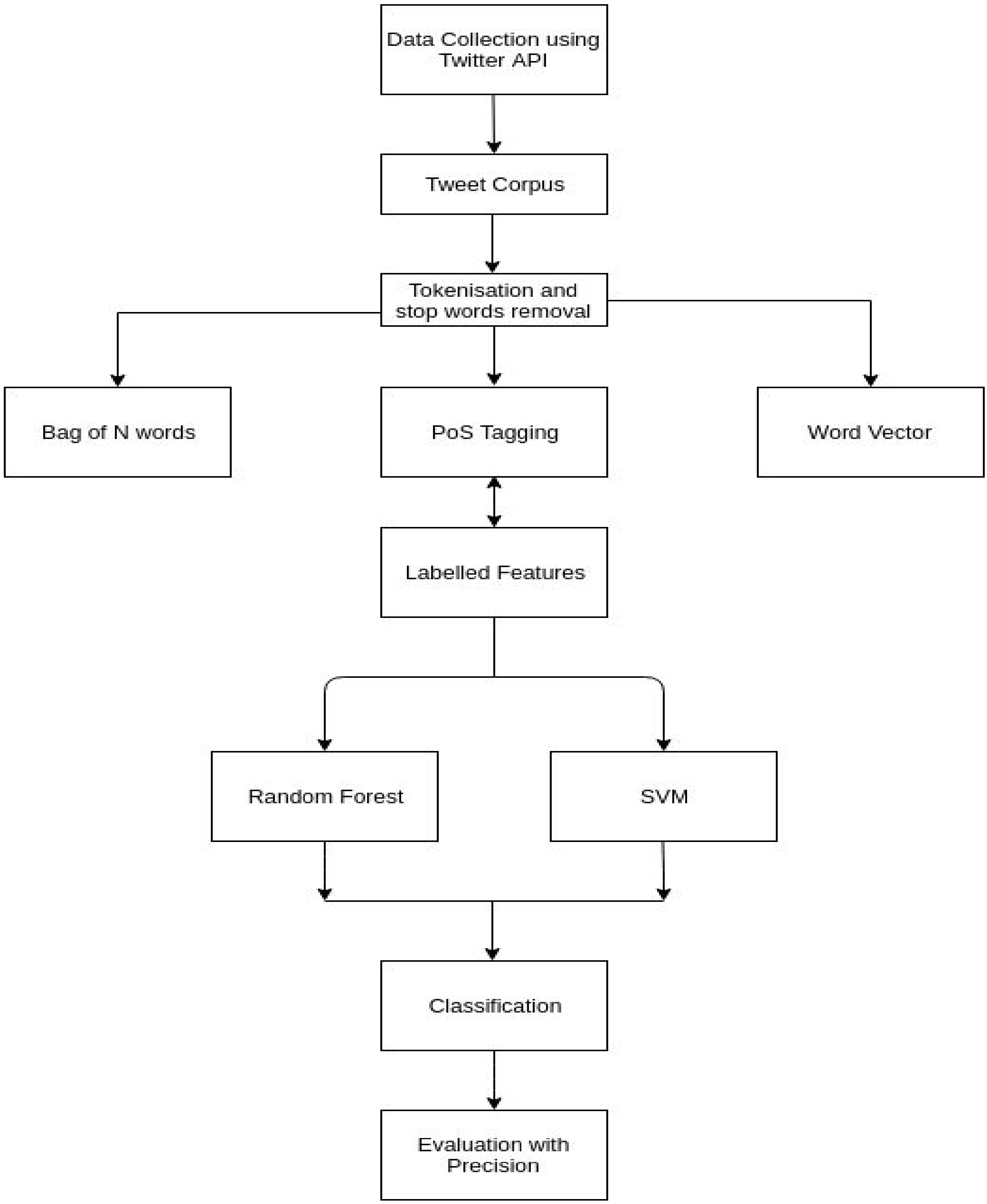


Figure 1 : Flow Diagram of the system

4.1 Data Collection :

- We collected tweets along with the user names that have personality tags such as #istj, #entp, etc.

- We then collect the recent tweets of users who tweeted about personality tags using their user ID collected during step 1.

- We compile a Twitter dataset with around 1,000 users by extracting and filtering all personality-related tweets on Twitter . The dataset contains not only the personality types, but also the most recent tweets for all the 1,000 users, which creates enormous opportunities to study the relationship between tweets and personality.

- We used “tweepy” package in python to do the data collection.

4.2 Feature Extraction

We design and implement three categories of linguistic features based on tweets, and further explore the correlations between each linguistic feature and each personality type. Several interesting findings can be observed here. For instance, extroverts tend to use hashtag and phrases like ”so proud”, ”so excited”, and ”can’t wait”. People who like to use emoticon are more likely to be Sensing and Feeling personality type.

4.2.1 Bag of N grams :

we use the most frequent 1000 unigram, bigram, trigram words and phrases from the all the tweets in our dataset. For 1-gram word, we remove stop words that don’t provide useful information about the language. The reason we only use the most frequent 1000 words and phrases is that data sparsity is a major problem in building language models. Using the most frequent 1000 words and phrases can effectively reduce the size of feature sets while still keep the valuable signals in these linguistic features.

Bag of n-grams model is used here to represents linguistic features of users based on n-gram models. By combining all the tweets for each user, we count number of occurrence for each frequent unigrams, bigrams, and trigrams and normalize by the number of unigrams, bigrams, and trigrams for each users. In this way, we construct a vector representation (1000 dimensions for unigram, bigram, and trigram respectively) for each user using language model. The bag of n-grams vector representations can be used to measure the correlations between n-gram words and phrases, and the personality traits.

4.2.2 PoS Tagging :

We adopt a POS tagger which has over 90% accuracy. Unlike, traditional POS tagger based on Penn Treebank, this Twitter POS tagger has 25 types of distinctive tags with some Twitter-specific tags such as hashtag, at-mention, discourse marker, URL, and emoticon.

We compute the distribution of each POS tag for users in our dataset. We compute the Pearson correlations between each POS tag and each personality trait. For example, in nominal and nominal+verbal category, common noun is a good indicator for personality. People who use common nouns more often in their language tend to be in Extroversion, Intuition, Thinking, or Judging type. Introverted people use more pronouns but less common nouns.

Interjection, which includes ”lol”, ”haha”, ”FTW”, ”yea”, is more likely to be used by people who are in Sensing and Perceiving type. Emoticon is more likely to be used by people who are in Sensing and Feeling type while numbers are more likely to be used by people who are in Sensing and Thinking type. Also, extroverted people are more likely to use hashtags. Those seemingly random online behaviors, such as use of hashtags, emoticons, and nouns can be somehow explained by the psychological traits of the users.

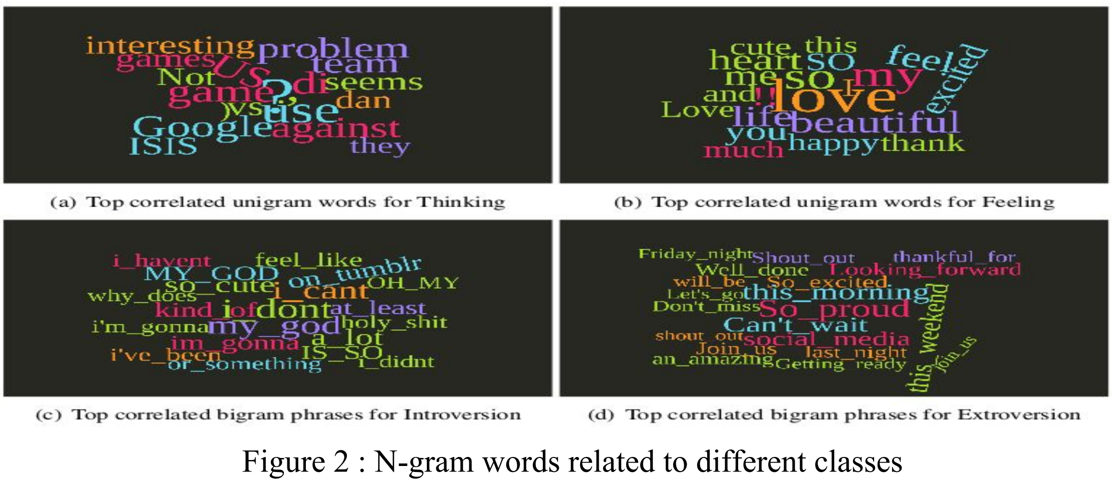
4.2.3 Word Vectors :

We want to explore the relationship between vector representations of words in tweets and the personality traits. In order to improve the generalization of the word vectors we use, we trained the model based on an external Twitter dataset which has hundreds of millions of
tweets. This gave us word vectors of 2,334,564 words and each word has a 500 dimension dis-
tributed representation of its semantic meanings.

Aiming to predict the personality traits, we need to compute a general representations of all tweets from a user based the word vectors. After removing stop words in all the tweets, we use two approaches to compute the textual representations of users listed as below.

• Average word vectors. We average all the vectors of all the word that is available in the tweets of a user to represent the vector representations of that user.

• Weighted average word vectors. Instead of average all the vectors of all the word equally, we weighted average the vectors of the words that is available in the tweets of a user according to the TF-IDF values. The weighted vector representation is then used to represent the vector representations of that user.



(a) Top correlated unigram words for Thinking
(b) Top correlated unigram words for Feeling
(c) Top correlated bigram phrases for Introversion
(d) Top correlated bigram phrases for Extroversion

Figure 2 : N-gram words related to different classes

4.4 Predicting Personality

We develop different prediction models using both individual features and combined features. For combined features, we concatenate features within or across categories to test the predictive power of our models. We used Random Forest and SVM to classify.

Support vector machines (SVMs) are a set of supervised learning methods used for classification, regression and outliers detection.

The advantages of support vector machines are:

- Effective in high dimensional spaces.
- Still effective in cases where number of dimensions is greater than the number of samples.
- Uses a subset of training points in the decision function (called support vectors), so it is also memory efficient.
- Versatile: different Kernel functions can be specified for the decision function. Common kernels are provided, but it is also possible to specify custom kernels

A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and use averaging to improve the predictive accuracy and control over-fitting. The sub-sample size is always the same as the original input sample size but the samples are drawn with replacement if bootstrap=True (default).

5. Performance

5.1 Performance of Individual Features and Feature Sets

The highest accuracy among individual features was achieved for 500 dimension vector representations of users based on the word vector. This is a remarkable performance given that those word vectors were unsupervisedly trained on a external Twitter dataset Likes and most of words are only weakly correlated with the personality traits. This result nicely illustrates the potential of predictions based on social media in general and languages in. Even the relatively weak signal aggregated over many observations (e.g. many words) results in a good prediction performance. Unigram, bigram, and trigram were also reasonably predictive of the personality traits. Beside, it is worth noting that the length of the n-gram feature also affect the prediction results. As trigram contains a sequences of three words, it introduces more noise into the feature and therefore performs worse in the prediction comparing to bigram and unigram.

POS tags give the lowest predictive accuracy among the three feature categories. POS tags convert all the tweets of a users into a distribution of 25 tags, much useful information
might be lost during this process which leads to lowest but moderate performance.

5.2 Performance of Combined Features :

In practice, features belonging to separate categories are often combined to maximize the model
performance. The model combining POS tag, Bag of n-grams, and word vector turns out to be very predictive of the participation. Removing word vector from the feature set significantly decreases the predictive power proving that the word vector is a very import feature that contains richer information about the personality traits than other two features (as shown in performance of individual features).

6. Conclusion

In this paper, we study the relationship between human language on Twitter and personality traits. Specifically, we want to know how linguistic features correlate with each personality trait and to what extent can we predict personality traits from language. We gather personality data from Myers-Briggs personality test which contains thinking, feeling, sensation, intuition, introversion, extroversion, judging and perceiving. Also, we collect 200 most recent tweets from users with personality values. We design three categories of feature, namely bag of n-grams, Twitter POS tags, and word vectors. Analysis of these features provide insights of language use for different personalities. For instances, extroverts tend to use hashtag and phrases like ”so proud”, ”so excited”, and ”can’t wait”. Moreover, we investigate the predictive power of individual features and combined features in our analysis. With the concatenation of all the features we extracted, we can predict the personality traits with an average 60%.

Social media is one of the most frequent destination for internet users. Inferring the personality traits of users in social media not only helps us understand their online behaviors, but also gives us the information to provide better personalized services and improve the product. However, predicting personality can also lead to privacy issue that expose the psychological details of online users to the public. These reasons make this study more important that it tries to understand personality traits from social media and explores the degree to which we can predict personality traits simply
using language on social media.

7. Future Work

While we were able to predict Extroversion, Introversion and Thinking, Feeling, the model was not as good at predicting Judging, Perception and Sensing, Intuition. To make the model better at these 2 categories as well, we hope to do the following in the near future:

1. Identify tags/characteristics that are attributes of each type in MBTI classification.

2. Create features to predict open categories and types outside of MBTI classification.

3. Use the combined value of these features together with a threshold to determine the MBTI categories indirectly.

8. References

1. Isabel Briggs Myers. 1962. The myers-briggs type indicator: Manual (1962).

2. Daniele Quercia, Michal Kosinski, David Stillwell, and Jon Crowcroft. 2011. Our twitter profiles, our selves: Predicting personality with twitter. In Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on, pages 180–185. IEEE.

3. Chris Sumner, Alison Byers, Rachel Boochever, and Gregory J Park. 2012. Predicting dark triad personality traits from twitter usage and a linguistic analysis of tweets. In Machine Learning and Applications (ICMLA), 2012 11th International Conference on, volume 2, pages 386–393. IEEE.

4. Ryan L Boyd, Steven R Wilson, James W Pennebaker, Michal Kosinski, David J Stillwell, and Rada Mihalcea. 2015. Values in words: Using language to evaluate and understand personal values. In Ninth International AAAI Conference on Web and Social Media.