Emotion Detection on Twitter Data

Sina Zamani

Advisors: Dr. Sauleh Etemadi, Hadi Sheikhi

April 2023

1 Introduction

Text classification methods can help us classify data into insightful categories. One of the main use cases of these classifiers is to determine several aspects of human cognition. As an important building block of human existence, we have to investigate human emotions which helps us in predicting future decisions and behaviors.

2 Data

2.1 Source

Social media monitoring becomes extremely helpful in terms of finding meaningful human behavioral patterns. People usually use these platforms to openly share their experiences and deep feelings.

Twitter is mostly a textual platform which gives us a huge corpus of human statements. Also it is known to be a place in which people broadcast their thoughts more frankly.

2.2 Language

Persian Twitter is the main context of this project but it is expected that with minor changes the model can be trained on non-persian tweets.

2.3 Formality

People post on twitter with a variety of different intentions. Some find it a friendly place to share their thoughts and some people may write a scientific thread discussing a special subject. So both formal and informal speech is used by tweeters.

2.4 Classifier

There is a variety of classifiers that have been proposed to define different human emotions. The one that we use, suggested by Paul Ekman, is one of the most cited ones and has six basic emotions including sadness, happiness, fear, anger, surprise, and disgust.

2.5 Motivation

As a psychology enthusiast, I find it interesting to find out how people behave in special circumstances and why so. This experiment becomes vastly useful when combined with other human cognitive aspects, helping us improve conscious behavior.

2.6 Assumptions

Aiming to predict emotional tone of any random tweet on twitter, we aren't taking any special time period, situation, or people category as assumption.

This can decrease our accuracy because the data is extremely general and doesn't follow any special rule. On the other hand, the results are less biased and can be interpreted as a more general understanding.

2.7 Collecting Data

Twitter has provided it's clients with an api to crawl this platform and extract tweets. Some necessary pre-processing will be performed on these data points including removing hashtags, URLs, and mentions.

People nowadays spend a huge portion of their time in social media posting new content. So there is no limitation in the quantity of data we can gather.

3 Challenges

3.1 Technical

- Data pre-processing
- Noisy data
- Data volume
- Imbalanced dataset
- Model selection

3.2 Cognitive

- Limited context
- Sarcasm and irony
- Biases
- Evolving/Dynamic language
- Human personality uniqueness