Feature Selection

Motivation

Why are ML-based things important for automatic analysis?

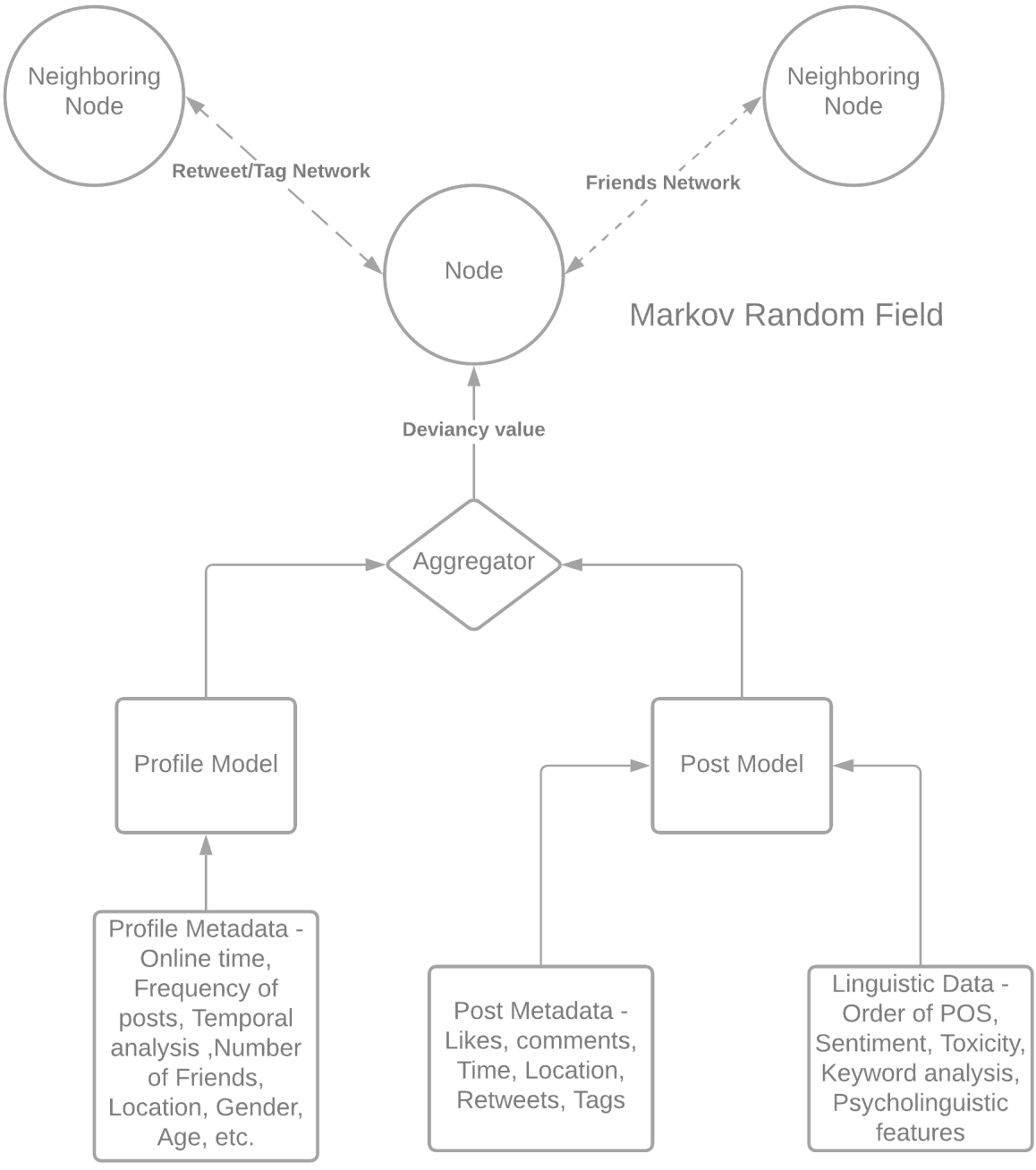
- Can find hidden patterns
- Can compute and discover more relations than a human
- Machine learning techniques identify high-quality solutions to mental health problems among Facebook users [link]
- The high volume of data, impossible to manually iterate [pastebin] (Notice the sheer amount of metadata for a single tweet)
- ML makes textual analysis much faster and more efficient than manual processing of texts.

Why is social network footprints more informative for this analysis?

- People post to express their feelings more on social media than real life nowadays as it’s not face-to-face
- In recent years, a number of studies have linked heavy social media use to an increased risk of depression.
- Overall, depression risk rose in tandem with time spent on social media. Compared with the lightest users (2 hours or less per day), the heaviest users (at least 5 hours per day) had three times higher depression risk. Meanwhile, that risk was two times higher among young adults who were active on social media around 3.5 to 5 hours per day.
- We find that social media contains useful signals for characterizing the onset of depression in individuals, as measured through a decrease in social activity, raised negative effect, highly clustered ego networks, heightened relational and medicinal concerns, and greater expression of religious involvement [link]

- Problematic SNS usage is significantly and positively related to depression and Neuroticism, while negatively associated with Agreeableness [link]
- There is a significant positive relationship between problematic SNS usage and depressive symptomatology
- O'Dea et al. [link] examined that Twitter is progressively researched as a method for recognizing psychological well-being status, including depression and suicidality in the population
- The researchers reported good performance in detecting depression by analyzing a variety of features such as lexical features including symptom lexicons (De Choudhury et al., 2013a; De Choudhury et al., 2013b; Coppersmith et al., 2014), syntactic features (Nambisan et al., 2015; Gkotsis et al., 2016), sentiment analysis (Wang et al., 2013; Preoţiuc-Pietro et al., 2015), or topic modeling (Resnik et al., 2015; Preoţiuc-Pietro et al., 2015)
- Language in social media activities is known to represent the current state of writers including their mental health. By analyzing the language used in social media, many researchers have discovered a way to identify depressed individuals.
- Many researchers have demonstrated that utilizing user-created content (UGC) accurately may help decide individuals' psychological wellness levels. For instance, Aldarwish and Ahmad [link] examined that the utilization of Social Network Sites (SNS) is expanding these days, particularly by the more youthful eras. Because the accessibility of SNS enables clients to express their interests, sentiments and offer day-by-day schedules [link, link].

Features



Structure of the Model Ensemble

- We use Belief Networks and Markov random field to propagate “Deviant” behavior.
- MRFs are a class of graphical models particularly suited for solving inference problems with uncertainty in observed data. MRFs are widely used in image

restoration problems wherein the observed variables are the intensities of each pixel in the image, while the inference problem is to identify high-level details such as objects or shapes. An MRF consists of an undirected graph, each node of which can be in any of a finite number of states. The state of a node is assumed to statistically depend only upon each of its neighbors, and independent of any other node in the graph. The dependency between a node and its neighbors is represented by a Propagation Matrix (ψ), where ψ(i, j) equals the probability of a node being in state j given that it has a neighbor in state i.

- The main motivation behind the usage of MRFs is that, some people have very low but niche footprint which maybe enough to conclude if a person is deviant but may escape/be misjudged by the model which can be caught using the propagation.

https://ermongroup.github.io/cs228-notes/representation/undirected/

- Here, social media accounts with very little profile and post metadata are considered uncertain and the "deviancy" values assigned to them are only propagated.
- Each node also stores the distribution of the features that contributed to the total Deviancy value hence addressing the Traceability concern.
- We can adapt an incremental version of MRF, which limits the propagation but keeps the difference within error margin; which can drastically increase the computational speed and hence make it suitable for Real Time Analysis.
- 2 network layers - Friends layer, Retweet and Tag layer

Profile and Post Metadata

- Metadata like location can also be an important factor in the analysis as certain places could be associated with a higher concentration of depressed individuals

- People who go through the same experiences in day to day life tend to have a similar outlook of the world, so exploring the friend network of a depressed individual could unlock more potential research
- Wolfradt and Doll (2001) suggested that gender is a major factor to consider when researching Internet use or use of Social Networking Sites
- Geographical location was researched as a factor mainly in terms of access to the internet and the “digital divide”, that is the unequal access to computers and the internet (Dewan & Riggins, 2005); urban areas are more likely to have internet access than rural areas, even though this fact is rapidly changing. A study on Norwegian adolescents (Johansson & Götestam, 2004) showed that the frequency of problematic Internet use was relatively higher in small cities and rural areas than in large cities (more than 150,000 inhabitants).
- Lastly, the number of posts and the temporal analysis of the activeness of an individual could be a key measure as well, as the activity tends to drop a lot or shoot up a lot seeking attention if one falls into depression

Linguistic Analysis

- We can use pre-existing vocabulary packages for deep and detailed analysis of the text like the LIWC.

  Psycholinguistic features LIWC is a psycholinguistic vocabulary package made by psychological analysts to perceive the different affective, intellectual, and etymological parts that lie on the user's verbal or written correspondence. It returns more than 70 different factors with a higher level of psycholinguistic features, for example,

    - Psychological process—effective process, social process, cognitive process, perceptual process, biological process, drives, time orientations, relativity, personal concerns
    - Linguistic process—word count, word/sentence, pronoun, personal pronoun, articles, prepositions, auxiliary verbs, adverbs, conjunctions, Negations
    - Other grammar—verbs, adjectives, comparisons, interrogatives, numbers, quantifiers.

These higher-level categories are also divided into subcategories such as

  - Biological processes—sexual, body, ingestion, and health
  - Affective processes—anxiety, anger, sadness, positive emotion, negative emotion
  - Time orientations—present, past, future
  - Social processes—family, friends, male, female
  - Perceptual processes—see, hear, feel.

- Identify Depressive symptoms using evidence keywords taken from a lexicon of nine groups of depressive symptoms in the Diagnostic and Statistical Manual of Mental Disorders (DSM-V)
- Analyze the sentiment of the posts as depressed people tend to have negative polarity in their posts
- Identify Ruminative thinking patterns as depressed people tend to have repetitive thoughts
- Looking at the POS (Part of Speech) level in their writings as their writing style tends to contain a different distribution of nouns, verbs, and adverbs and the complexity of sentences (Gkotsis et al., 2016)