Real-time Social Media Sentiment Analysis

Statistical Learning and Data Analytics

Project Proposal

Can Kocagil

Barış Kıcıman

EEE 485/585

Department of Electric & Electronics Engineering

Bilkent University

Ankara, Turkey

12.10.2021

1. Introduction

Social media is a place where distributed social agents express their opinions, daily interactions, emotions, and feelings by posting multiple contents, tweets, images, video etc., such as on Facebook, Twitter and Instagram. With the massive growth in social web services, and media, these digital platforms reshapes the consumers preferences, and daily lives in a inevitable way. As the consumers daily preferences are ever-changing, the value-driven enterprises moved their digital marketing agencies to social medias, especially in Facebook and Twitter. Hence, these ever-growing social data in any kind contains natural human-behavior based information that facilitates the autonomous decision making. To extract behavioral information from complex and distributed social media data, sentiment analysis literature is emerged and powered by statistical learning theories and deep learning. By acknowledging the social-economical context, we determined to develop behavior-aware machine learning algorithms, specialized for understanding the social text data by extracting behavior-oriented sentiments.

2. Proposal

As the project consist of multiple layers, we'll introduce different learning algorithms for each phase of the project, by fixing the training data but varying the testing data. As the sentiment analysis task is supervised, we need to have labels for the social text data that represents the sentiments annotated by human specialists or autonomous annotation system. To capture social media dynamics, and stochasticity driven by human psychology, we looked for highly generic social text data which hopefully scalable and generalizable to real-life applications in business context. We found that Sentiment140 [1] dataset is appropriate for our needs, and the scope of the project, as it consist of tweets and sentiments of a brands & products, that are highly scalable for brand needs, managements, operations, and pollings. The dataset consist of 1.6 million tweets with their corresponding sentiments, that are broadcasted between $[0, 4]$, being the polarity of the tweet 0 = negative, 2 = neutral, and 4 = positive. By the inference time, we'll fetch real-time tweet data from Twitter, and produce predictions in a real-time. As the social tweets are highly unstructured, having a large corpus and stochastic by nature, the curse of dimensionality, the extraction of semantic information, and being real-time are our main challenges we'll face with, that constraints our computational comfort zone. From the machine learning perspective, we'll design & develop classifier-based learning algorithms, as of start, we'll introduce Naive Bayes, to construct our baseline model. Then, we'll move to more advanced machine learning algorithms such as SVM. Then, we'll construct Deep Neural Network (DNN) architectures with both embedding & linear layers, to capture semantic information lying the tweet data, that hopefully extract behavioral sentiments in a accurate manner.

3. Conclusion

By recognizing the significance of autonomous analysis of social media data that represents human preferences, daily interactions, and behavioral cognition, we'll design cognitive machine learning algorithms, from Naive Bayes to Deep Neural Networks, to capture the social-economical dynamics by extracting sentimental information. As the cognitive machine learning models are fueled by behavior-oriented data, we'll determined to utilize Sentiment140 [1] dataset that consist of 1.6 Million social tweets from brands & products, to train our models. By the prediction time, we'll fetch real-time tweet data to analyze our models real-time performance, as it super-significant aspect of any statistical model in the context of learning theories.

REFERENCES

[1] Alec Go, Richa Bhayani, and Lei Huang. "Twitter sentiment classification using distant supervision". In: *CS224N project report, Stanford* 1.12 (2009), p. 2009.