CAN MICROBLOGS PREDICT MUSIC CHARTS?

An Analysis of the Relationship between #nowplaying Tweets and Music Charts

Eva Zangerle, Martin Pichl, Benedikt Hupfauf, Günther Specht
Department of Computer Science, University of Innsbruck, Austria

MOTIVATION

People tweet about music they are listening to in #nowplaying tweets. In this work, we investigate whether this information can contribute to predicting future charts. Particularly, we look at the following three core questions:

- How are #nowplaying tweets and the Billboard Hot 100 temporally related?
- How can Twitter data be exploited for predicting music charts?
- To which extent do #nowplaying-tweets resemble the Billboard Hot 100?

We aim to perform these analyses by modelling Twitter and charts data as time series

DATA

Twitter: #nowplaying tweets of 2014 and 2015 (111,260,925 tweets)

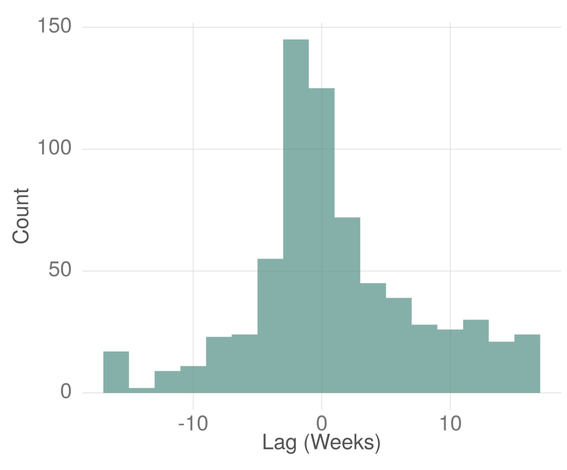
Charts: Billboard Hot 100 of 2014 and 2015 (886 tracks)

To be able to compare these data sets, we compute the overlap of tweets and charts by matching tweet text and artist and track of Billboard data:

- Track: matches if it is contained in the tweet.
- Artist: matches if more than 50% of the tokens are contained in the tweet text as there exists various formats as e.g., Michael Jackson vs. Jackson, Michael; featuring vs. ft. vs. &, etc.

TEMPORAL ASPECTS

Cross-correlation analysis of Twitter and Billboard time series to compute lag between time series representing Twitter and Billboard data.



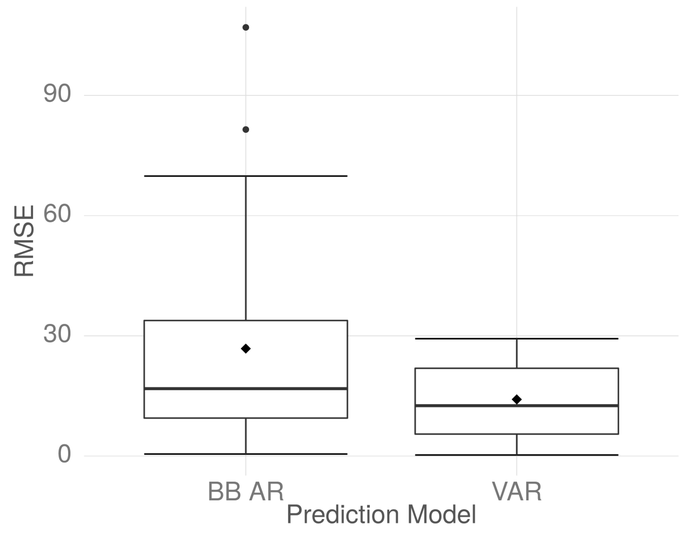
	Min	Q1	Med	Mean	Q3	Max
All	-17.0	-2.0	0.0	1.47	5.0	17.0
Twitter first	-17.0	-2.0	0.0	0.97	4.0	17.0

Lag (in Weeks): Five-Number Summary

CHARTS PREDICTION

3 prediction models:

- Autoregression based on Billboard time series (BB)
- Extract the lag from cross-correlation analysis, shift base and and compute autoregression based on the shifted difference between Twitter and Billboard (T) .
- Multivariate model based on Twitter and Billboard time series (V).



	Min	Q1	Med	Mean	Q3	Max
T	16.3	84.5	116.1	148.6	178.3	388.4
BB	0.51	9.5	16.8	26.8	33.8	107.0
V	0.27	5.5	12.6	14.1	21.9	29.3

RMSE Five-Number Summary

MAIN FINDINGS

- From a temporal perspective, there is a positive lag for 48% of all tracks. 11% do not feature a lag.
- 41% of all tracks feature negative lag and would hence allow for a prediction.
- The multivariate model based on Twitter and Billboard charts data significantly reduces the RMSE ($p < 0.05$; Mann-Whitney) when compared to the Billboard-based model.
- Variance of multivariate model shows significantly lower variance of RMSE than the Billboard-based model ($p < 0.05$; Levene).
- Mild correlation for track playcounts on Twitter and Billboard charts ($p < 0.01$; Pearson).

CONTACT

Eva Zangerle
http://www.evazangerle.at
http://dbis-informatik.uibk.ac.at
eva.zangerle@uibk.ac.at