Buzz prediction on Twitter

This data-set contains examples of buzz events from Twitter about some topics.

Objective

We want to predict if a topic could gain popularity in a social network.

Predicted attribute :

"Annotation" (Buzz). This attribute is binary: 1 meaning "buzz observed", 0 meaning "no buzz observed". It is stored is the rightmost column.

Relevant Information Paragraph:

Each instance covers seven days of observation for a specific topic (eg. overclocking...). Considering the couple day following this initial observation; If there is at least 500 additional active discussions by day (on average, with respect to the initial observation) then, the predicted attribute Buzz is True.

Observations are Independent and identically distributed.

Number of Instances: 126635

Number of Attributes: 77 + class attribute

Time representation: Each instance is described by 77 features, those describe the evolution of 11 "primary features" through time. Hence each feature name is postfixed with the relative time of observation. For instance, the value of the feature "Nb_Active_Discussion" at time t is given in "Nb_Active_Discussion_t".

Attributes

-- Number of Created Discussions (NCD) (columns [0,6])

-- Type : Numeric, integers only

-- Description : This feature measures the number of discussions created

at time step t and involving the instance's topic.

-- Columns : From column 0 (NCD at relative time 0) to column 6 (NCD at

relative time 6)

-- Abbreviations : NCD_0, NCD_1, NCD_2, NCD_3, NCD_4, NCD_5, NCD_6

-- Author Increase (AI) (columns [7,13])

-- Type : Numeric, integers only

-- Description : This feature measures the number of new authors interacting on

the instance's topic at time t (i.e. its popularity)

-- Columns : From column 7 (AI at relative time 0) to column 13 (AI at relative time 6)

-- Abbreviations : AI_0, AI_1, AI_2, AI_3, AI_4, AI_5, AI_6

-- Attention Level (measured with number of authors) (AS(NA))

(columns [14,20])

-- Type : Numeric, real in [0,1]

-- Description : This feature is a measure of the attention payed to a the instance's topic on a social media.

-- Columns : From column 14 (AS(NA) at relative time 0) to column 20 (AS(NA) at relative time 6)

-- Abbreviations : AS(NA)_0, AS(NA)_1, AS(NA)_2, AS(NA)_3, AS(NA)_4, AS(NA)_5, AS(NA)_6

-- Burstiness Level (BL) (columns [21,27])

-- Type : Numeric, defined on [0,1]

-- Description : The burstiness level for a topic z at a time t is defined as the ratio of ncd and nad

-- Columns : From column 21 (BL at relative time 0) to column 27 (BL at relative time 6)

-- Abbreviations : BL_0, BL_1, BL_2, BL_3, BL_4, BL_5, BL_6

-- Number of Atomic Containers (NAC) (columns [28,34])

-- Type : Numeric, integer

-- Description : This feature measures the total number of atomic containers generated through the whole social media on the instance's topic until time t.

-- Columns : From column 28 (NAC at relative time 0) to column 34 (NAC at relative time 6)

-- Abbreviations : NAC_0, NAC_1, NAC_2, NAC_3, NAC_4, NAC_5, NAC_6

-- Attention Level (measured with number of contributions) (AS(NAC))

(columns [35,41])

-- Type : Numeric, real in [0,1]

-- Description : This feature is a measure of the attention payed to a

the instance's topic on a social media.

-- Columns : From column 35 (AS(NA) at relative time 0) to column 42

(AS(NAC) at relative time 6)

-- Abbreviations : AS(NAC)_0, AS(NAC)_1, AS(NAC)_2, AS(NAC)_3, AS(NAC)_4,

AS(NAC)_5, AS(NAC)_6

-- Contribution Sparseness (CS) (columns [42,48])

-- Type : Numeric, real in [0,1]

-- Description : This feature is a measure of spreading of contributions

over discussion for the instance's topic at time t.

-- Columns : From column 42 (CS at relative time 0) to column 48

(CS at relative time 6)

-- Abbreviations : CS_0, CS_1, CS_2, CS_3, CS_4, CS_5, CS_6

-- Author Interaction (AT) (columns [49,55])

-- Type : Numeric, integer.

-- Description : This feature measures the average number of authors

interacting on the instance's topic within a discussion.

-- Columns : From column 49 (AT at relative time 0) to column 55

(AT at relative time 6)

-- Abbreviations : AT_0, AT_1, AT_2, AT_3, AT_4, AT_5, AT_6

-- Number of Authors (NA) (columns [56,62])

-- Type : Numeric, integer.

-- Description : This feature measures the number of authors interacting

on the instance's topic at time t.

-- Columns : From column 49 (NA at relative time 0) to column 55 (NA at

relative time 6)

-- Abbreviations : NA_0, NA_1, NA_2, NA_3, NA_4, NA_5, NA_6

-- Average Discussions Length (ADL) (columns [63,69])

-- Type : Numeric, real.

-- Description : This feature directly measures the average length of a

discussion belonging to the instance's topic.

-- Columns : From column 63 (ADL at relative time 0) to column 69 (ADL at relative time 6)

-- Abbreviations : ADL_0, ADL_1, ADL_2, ADL_3, ADL_4, ADL_5, ADL_6

-- Average Discussions Length (NAD) (columns [70,76])

-- Type : Numeric, integer.

-- Description : This features measures the number of discussions involving the instance's topic until time t.

-- Columns : From column 70 (NAD at relative time 0) to column 76 (NAD at relative time 6)

-- Abbreviations : NAD_0, NAD_1, NAD_2, NAD_3, NAD_4, NAD_5, NAD_6

-- Annotation (column 77)

-- Type : Numeric, integer: 0 or 1

-- Description : class attribute

-- Columns : 77

-- Buzz = 1

-- Non Buzz = 0