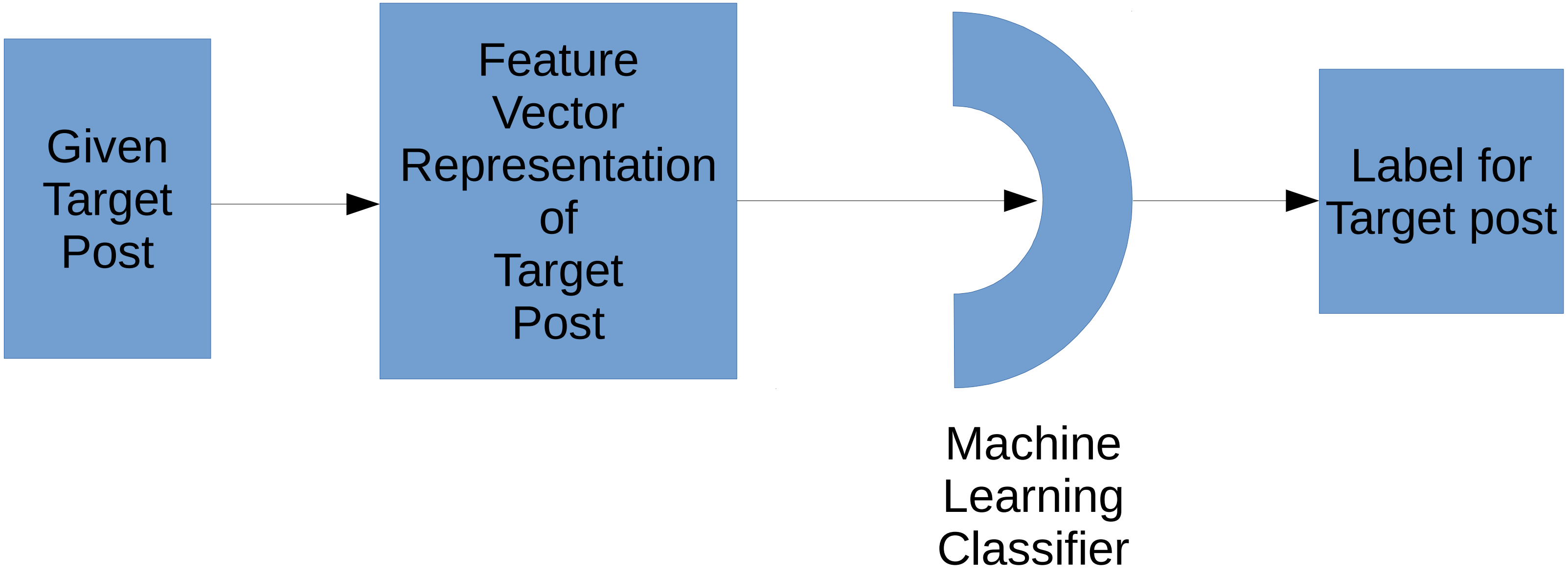
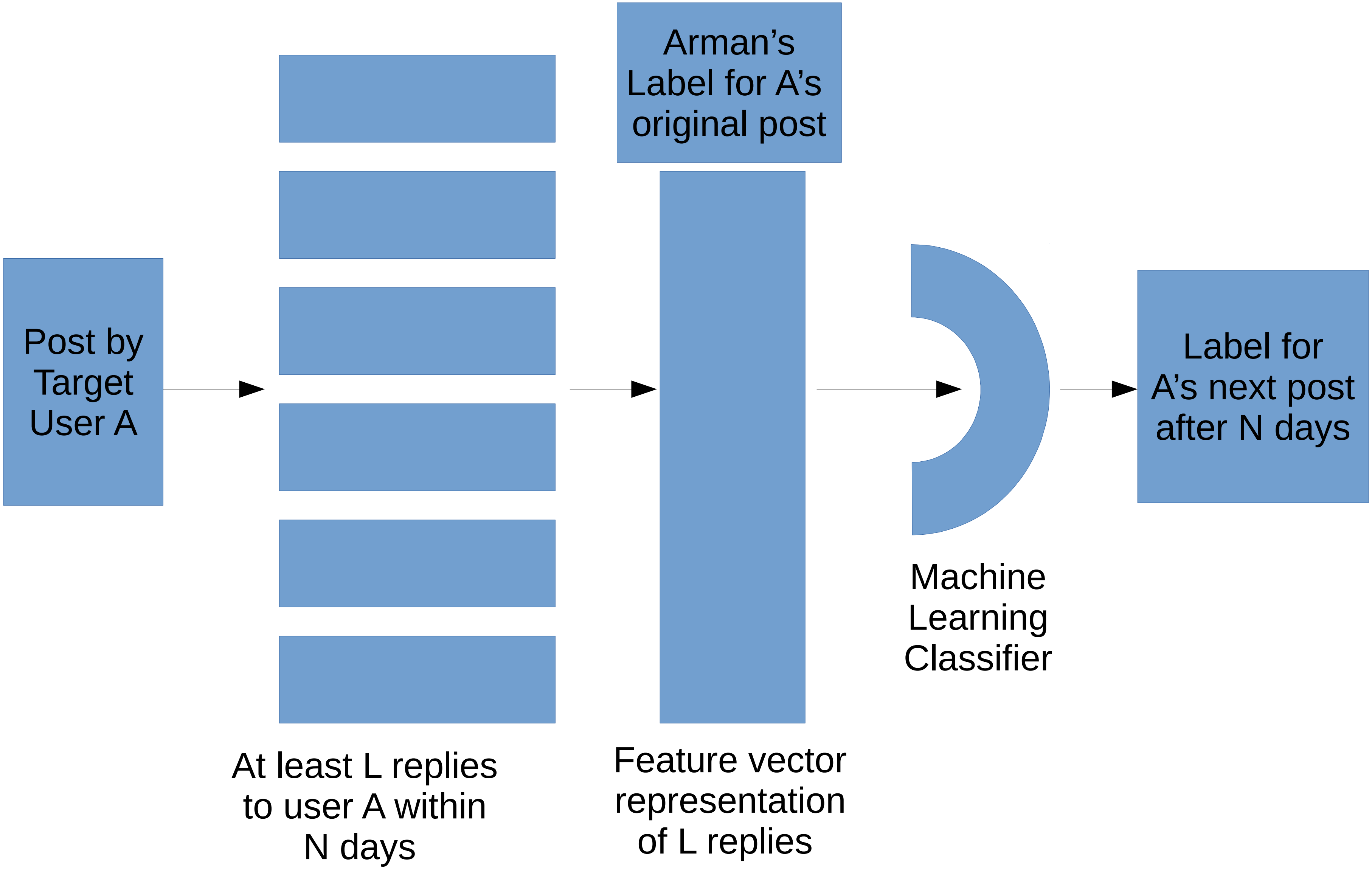
Old Task


Given Target Post
Feature Vector Representation of Target Post
Machine Learning Classifier
Label for Target post

New Task



Post by Target User A

At least L replies to user A within N days

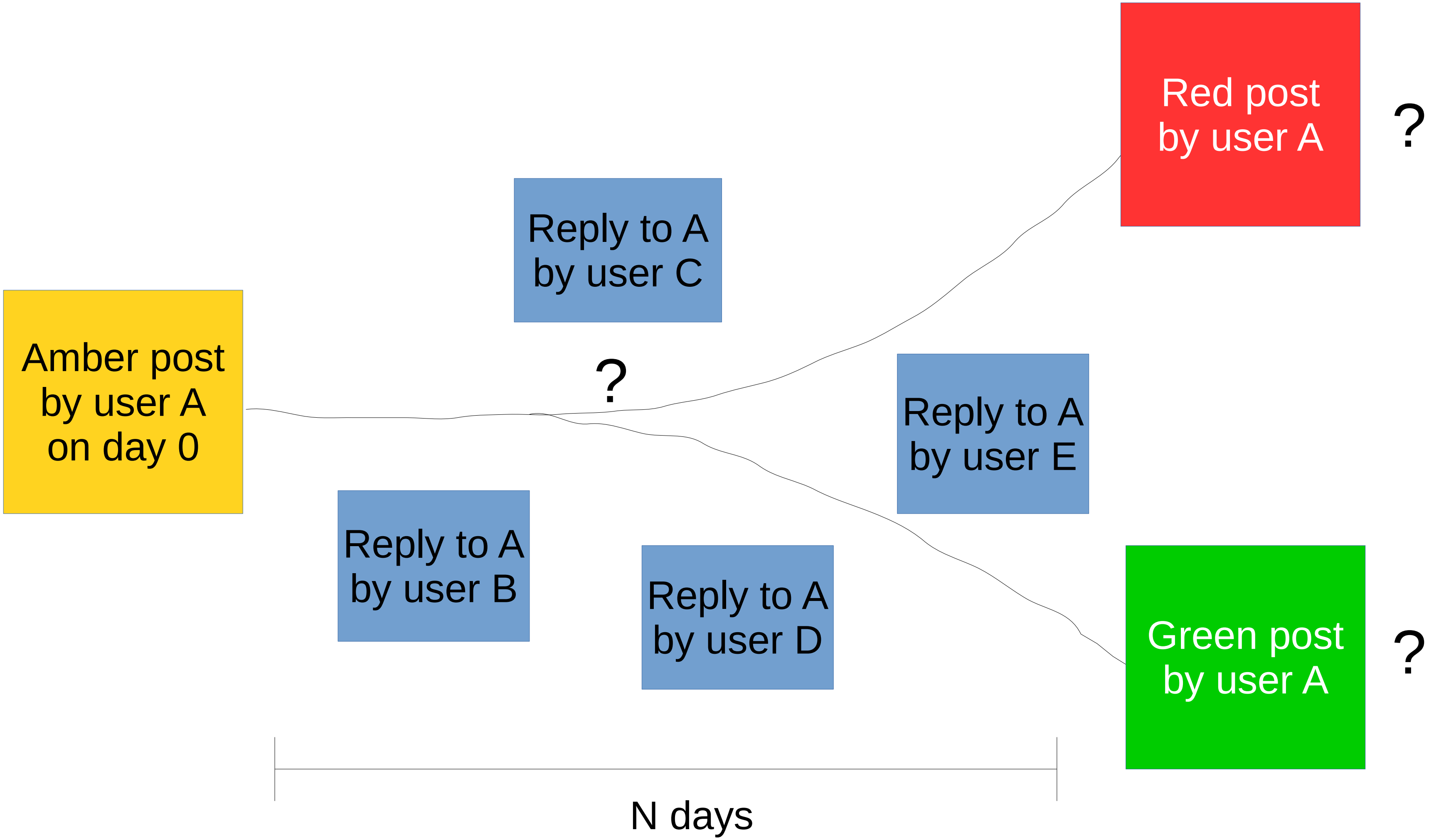
Arman’s Label for A’s original post

Feature vector representation of L replies

Machine Learning Classifier

Label for A’s next post after N days

New Task in Another Picture



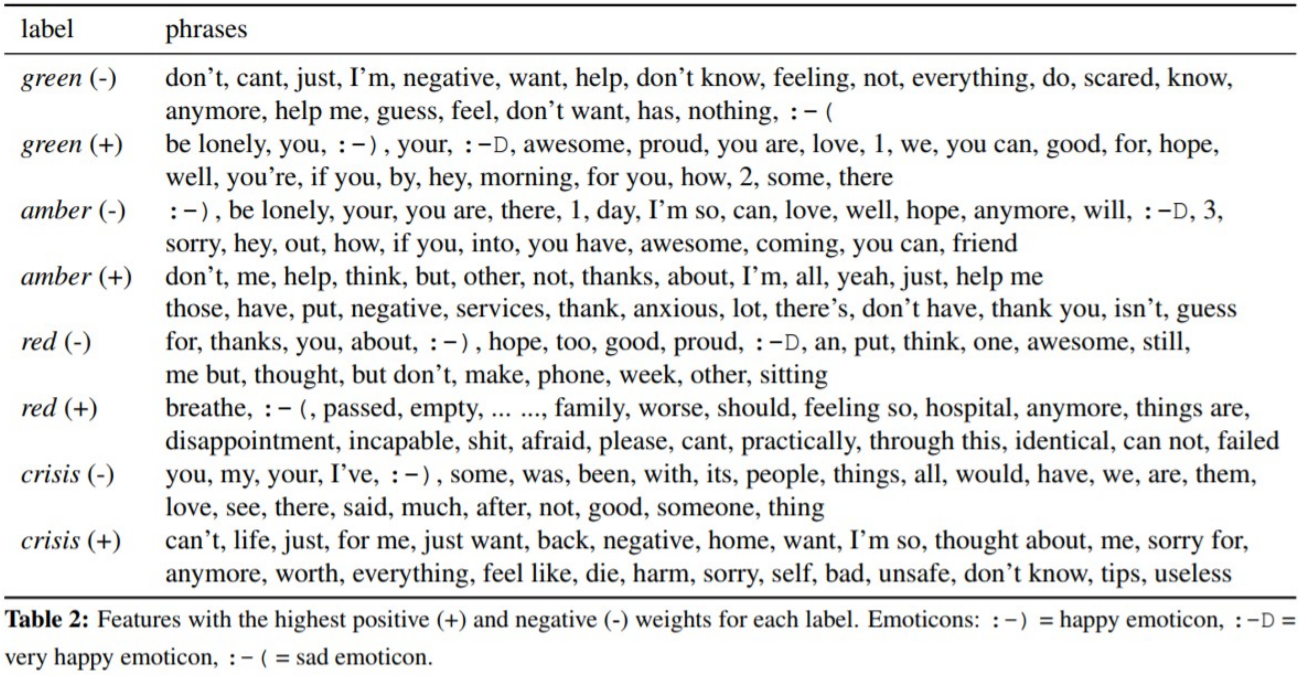
New Task in a Third Picture

[timestamp]	[id]	[login]	[text]	[users_quoted]	[Arman_label]
20160801	145	A	……	none	amber
⋮	⋮	⋮	⋮	⋮	⋮
20160802	171	B	…….	A	green
20160802	172	C	….	A	green
⋮	⋮	⋮	⋮	⋮	⋮
20160804	201	A	…….	none	green

Create feature vector from this

Try to predict this label

“Interpretable Model”



label	phrases
green (-)	don’t, cant, just, I’m, negative, want, help, don’t know, feeling, not, everything, do, scared, know, anymore, help me, guess, feel, don’t want, has, nothing, :-(
green (+)	be lonely, you, :-) , your, :-D, awesome, proud, you are, love, 1, we, you can, good, for, hope, well, you’re, if you, by, hey, morning, for you, how, 2, some, there
amber (-)	:-) , be lonely, your, you are, there, 1, day, I’m so, can, love, well, hope, anymore, will, :-D, 3, sorry, hey, out, how, if you, into, you have, awesome, coming, you can, friend
amber (+)	don’t, me, help, think, but, other, not, thanks, about, I’m, all, yeah, just, help me those, have, put, negative, services, thank, anxious, lot, there’s, don’t have, thank you, isn’t, guess
red (-)	for, thanks, you, about, :-) , hope, too, good, proud, :-D, an, put, think, one, awesome, still, me but, thought, but don’t, make, phone, week, other, sitting
red (+)	breathe, :-(, passed, empty, ..., family, worse, should, feeling so, hospital, anymore, things are, disappointment, incapable, shit, afraid, please, cant, practically, through this, identical, can not, failed
crisis (-)	you, my, your, I’ve, :-) , some, was, been, with, its, people, things, all, would, have, we, are, them, love, see, there, said, much, after, not, good, someone, thing
crisis (+)	can’t, life, just, for me, just want, back, negative, home, want, I’m so, thought about, me, sorry for, anymore, worth, everything, feel like, die, harm, sorry, self, bad, unsafe, don’t know, tips, useless

Table 2: Features with the highest positive (+) and negative (-) weights for each label. Emoticons: :-) = happy emoticon, :-D = very happy emoticon, :-(= sad emoticon.

“The ranking of features by their respective squared weights can be interpreted as metric of feature relevance (Guyon et al., 2002). High weights (their squared value to take negative weights into account) influence the output of the decision function by tendency more than low weights.”

Using Linear Classifiers for the Automatic Triage of Posts in the 2016 CLPsych Shared Task - Juri Opitz

Questions/Comments

- How much do L and N parameters matter? For small N, we cannot have a large L. For large N, we lose relevancy: a reply on day 2 probably doesn't affect a user's state on day 30… Maybe N = 7 is a good starting point?

- To be interpretable, many features are not helpful even if they are predictive, e.g. kudos. What about features like sentiment, emotion, or topic? Obviously specific words/phrases are best.

- Many reasons why we might find nothing…

  1. Offline events affect a user's state more than ReachOut.com activity.

  2. Survivorship bias: Users only continue activity on ReachOut.com if they are finding it helpful.

  3. Maybe everybody in a distressed state gets better over time for no apparent reason. Maybe simply writing about feelings is helpful, and replies have no effect at all.