

Economically-Efficient Sentiment Stream Analysis

Roberto Loureco Jr., Adriano Veloso, Adriano Pereira
Wagner Meira Jr., Renato Ferreira, Srini Parthasarathy

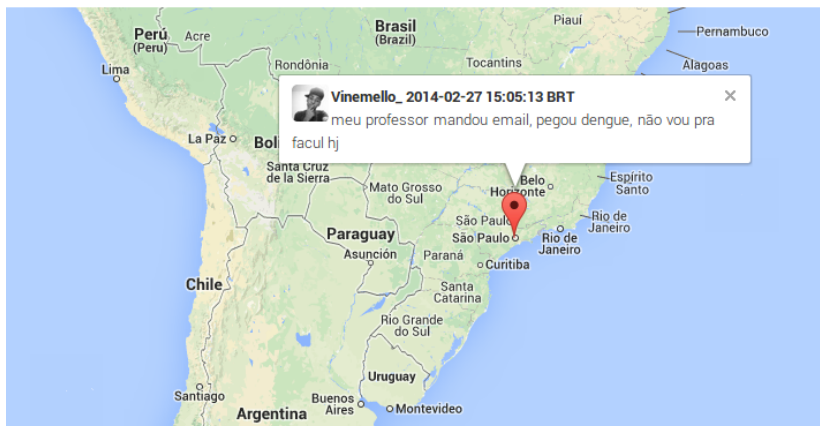
Computer Science Dept - UFMG - Brazil
Dept. Computer Science and Engineering - OSU

Sentiment Analysis

- Definition
 - Automatically extraction of opinions, sentiments, attitudes, and emotions expressed in text
- Motivation
 - It allows us to track products, brands and people to determine whether they are viewed positively or negatively.
- Problem
 - Content is created almost at the same time the event is happening in the real world.
 - **Sentiment streams.**

Sentiment Streams and Outbreaks

Dengue Fever



Sentiment Streams and Advertising

Superbowl 2013



Matt Hannaford @mhannaford

15h

Did Mercedes-Benz not pay the electric bill? #superbowl

Retweeted by Audi

[Collapse](#) [Reply](#) [Retweet](#) [Favorite](#) [More](#)

423

RETWEETS

102

FAVORITES



8:38 PM - 3 Feb 13 · [Details](#)

Sentiment Streams and Advertising

Superbowl 2013



Matt Hannaford @mhannaford

15h

Did Mercedes-Benz not pay the electric bill? #superbowl

Retweeted by Audi

[Collapse](#) [Reply](#) [Retweet](#) [Favorite](#) [More](#)

423

RETWEETS

102

FAVORITES



8:38 PM - 3 Feb 13 · Details



Audi @Audi

15h

Sending some LEDs to the @MBUSA Superdome right now...

[Collapse](#) [Reply](#) [Retweet](#) [Favorite](#) [More](#)

9,397

RETWEETS

2,980

FAVORITES



8:40 PM - 3 Feb 13 · Details

Sentiment Streams and Advertising

and yesterday ... Brazil 1 × Germany 7



PlayStation Brasil @PlayStation_BR · 2 h

#SeFosseNoPLAY era apertar o Reset e começar outra!

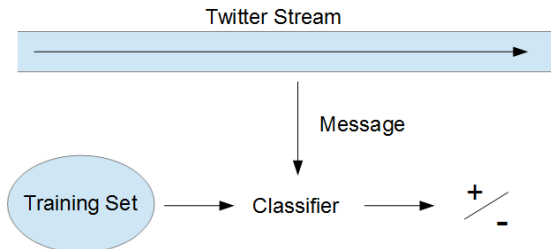
#BRA 🇧🇷 vs **#GER** 🇩🇪 pic.twitter.com/wMtwpsGNFF

Just hit the reset button and start again

#BRA 🇧🇷 vs **#GER** 🇩🇪

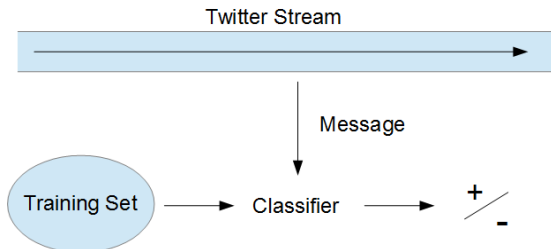
Classifying Sentiment Streams

- Classifiers may be used to distinguish sentiments in the text.



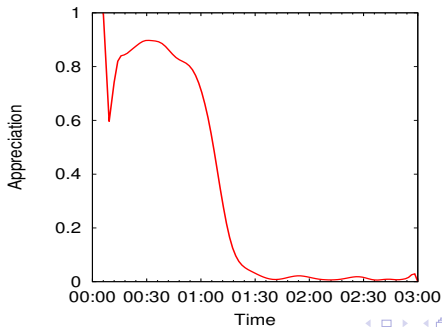
Classifying Sentiment Streams

- Classifiers may be used to distinguish sentiments in the text.

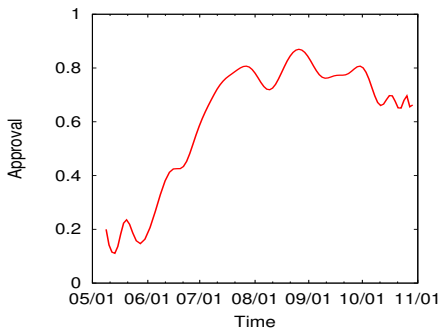


- Sentiments may change with time.

Sports (WC 2010)

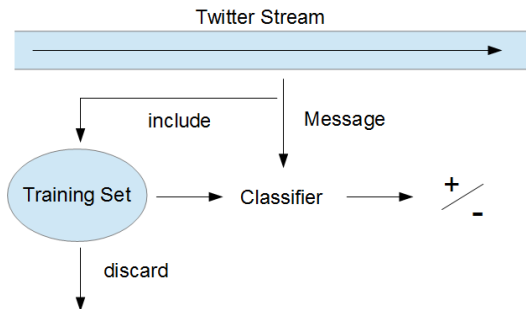


Elections (Brazil 2010)



Classifying Sentiment Streams

- Effective classification requires:
 - Updating the training-set to mitigate drifts.
 - Updating the classifier accordingly.



Research Questions

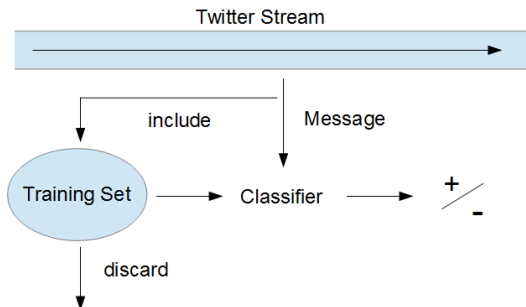
- 1 Effort:
 - How to reduce labeling effort?
- 2 Accuracy:
 - How to select messages to be included and discarded?

Research Questions

- ① Effort:
 - How to reduce labeling effort?
- ② Accuracy:
 - How to select messages to be included and discarded?

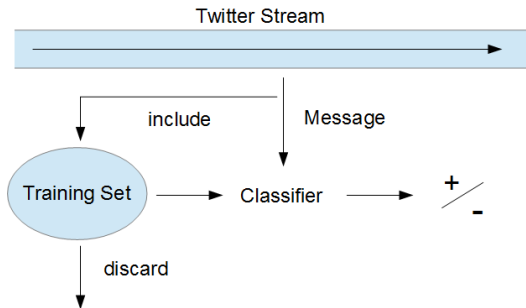
Batch Mode

- Instead of updating the training set on an instance-basis
 - Wait for a batch of b messages.
 - A similarity parameter δ controls the messages in the batch that must be labeled.



Batch Mode

- Instead of updating the training set on an instance-basis
 - Trade-off between batch size and accuracy.
 - The larger is the batch, the lower is the accuracy.

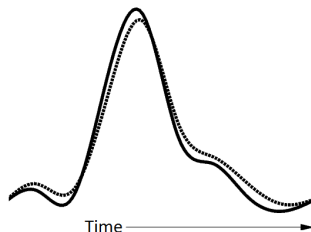
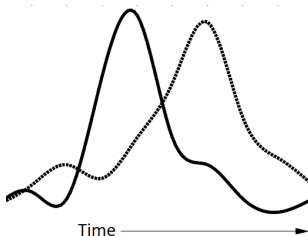


Research Questions

- ① Effort:
 - How to reduce labeling effort?
- ② Accuracy:
 - How to select messages to be included and discarded?

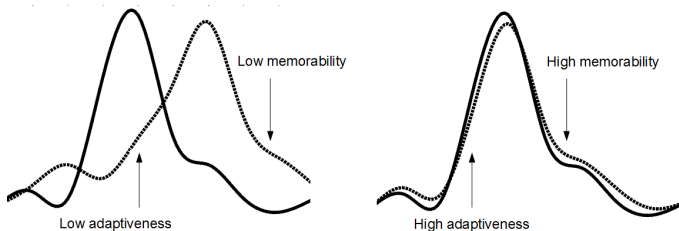
Dealing with Drifts

- Intuition:
 - We must gather relevant data as soon as possible.



Dealing with Drifts

- Two properties are necessary in order to produce classifiers that are robust to drifts:
 - Adaptiveness:
 - The ability to adapt itself to drifts.
 - Memorability:
 - The ability to recover itself from drifts.

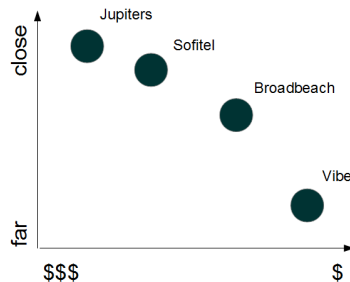


Dealing with Drifts

- Two properties are necessary in order to produce classifiers that are robust to drifts:
 - Adaptiveness:
 - The ability to adapt itself to drifts.
 - The training-set must contain fresh messages.
 - Memorability:
 - The ability to recover itself from drifts.
 - The training-set must contain messages belonging to pre-drift distributions.
- Improving both properties simultaneously may lead to a conflict-objective problem.
 - Improve adaptiveness may hurt memorability, and vice-versa.

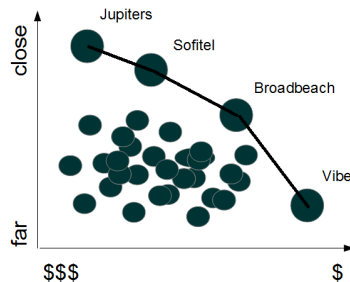
Pareto Efficiency

Hotels close to the Convention Center.



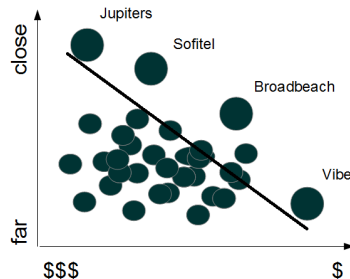
Pareto Efficiency

Pareto frontier.



Compensation — Kaldor-Hicks Principle

Region of compensation.

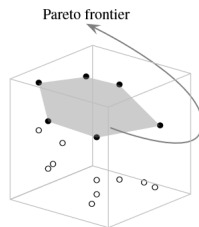
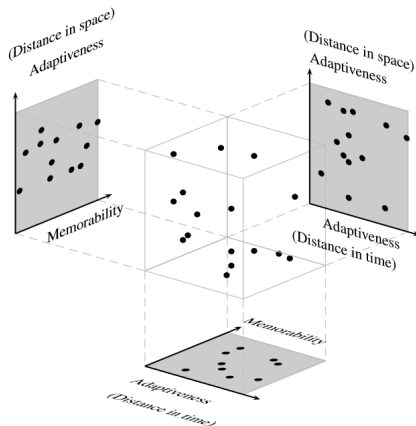


Utility Measures

- Distance in space:
 - $U_s(t_j) = \frac{|\mathcal{R}(t_n) \cap \mathcal{R}(t_j)|}{|\mathcal{R}(t_n)|}$
- Distance in time:
 - $U_t(t_j) = \frac{\gamma(t_j)}{\gamma(t_n)}$.
 - $\gamma(t_j)$ returns the time in which message t_j arrived.
- Random permutation of messages:
 - $U_r(t_j) = \frac{\alpha(t_j)}{|\mathcal{D}_n|}$
 - $\alpha(t_j)$ returns the position of t_j in the shuffle.
 - \mathcal{D}_n is the training set at time step n .

Utility Measures

- ① At each time step n :
 - ① Place candidate messages in the utility space.
 - ② Select messages in the Pareto frontier.

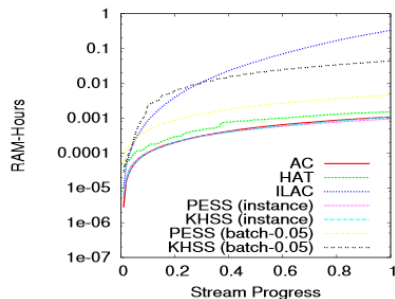
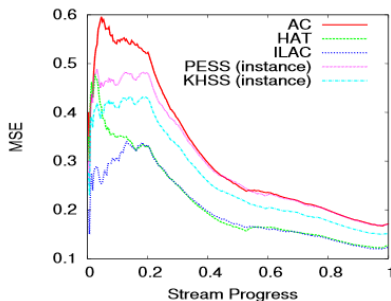


Evaluation

- Measures used:
 - Mean Square Error.
 - RAM-Hours:
 - A GB of RAM deployed for 1 hour execution.
- Labeling Effort:
 - Different batch sizes and δ values.
- Three datasets:
 - Brazilian elections 2010.
 - World Cup 2010.
 - Person of the Year 2010 (Assange vs. Zuckerberg)
- Baselines:
 - AC — Active Classifiers (KDD 2011)
 - HAT — Hoeffding Adaptive Trees (JMLR 2011)
 - ILAC — Incremental Lazy Classifiers (SIGIR 2011)

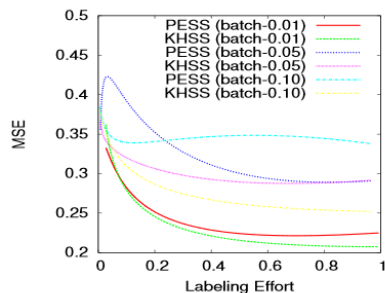
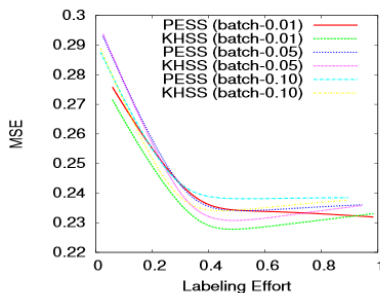
Evaluation

- MSE and RAM-Hours



Evaluation

- MSE and Labeling Effort



Conclusions

- Sentiment analysis on Twitter streams.
 - Limited computing and training resources.
 - Sentiment drifts.
- Efficiency and accuracy.
 - Incremental classifiers.
 - Pareto efficiency and compensation principle.
- Our results.
 - 50
- Future work includes:
 - Other utility measures.
 - Other application scenarios.

Thank you!

adrianov@dcc.ufmg.br