Dear SBBD 2011 chairs,

In this document we report a summary of the improvements to our paper "Exploiting Temporal Locality to Determine User Bias in Microblogging Platforms". We thank reviewers for their insightful reviews. In addition to minor corrections (i.e., typos and grammar errors) and an overall review of the text, the following modifications have been performed:

1 Review 1

It seems to me too strong to affirm that the proposed algorithm learns something. In fact, it is already based on a bias that messages issued shortly in time (in a range to be defined) follow the same bias. That is to say that, the algorithm is prepared to compute the bias of other users based on those with bias already known (assuming these users exist). based on several parameters. None of them are learned.

Actually, the only parameter of the algoritm is interval size; we included a new Section (4.2) which shows the impact of this parameter on results. We do not consider minimum interval size as a parameter because we have shown, in Section 4.3, that considering all intervals, or discarding the very low-volume ones, generates better results.

We think that our algorithm is a supervised learning algorithm because it uses pre-labeled information to predict unknown information. Our goal in this paper was to show that even a simple approach is able to capture meaningful patterns from temporal data in microblogs; that is why we didn't work on more sophisticated algorithms for now. We have added a sentence in the Conclusions saying that, besides showing that it is possible to use temporal information to predict user inclinations, our work is useful to motivate researchers to use the temporal dimension to build and design applications that use it to customize and enhance user experience in social media. We think this is an important aspect of our work that was not highlighted in our original submission. For example, a possible application is to design an real-time recommendation algorithm that recommends content in real-time, or recommends users that tend to post at the same time.

The authors say that they have "the assumption that bias does not happen randomly but usually occur in reaction to particular events". It is also said that "we argue the, in some scenarios, We can rely on when ..." What are these scenarios/circumstances? The case studies have similar nature like a political debate and matches. Does it requires the identification of an event or reference or it would be possible to apply the method in any type of stream of messages?

The assumption that enables our algorithm to work is that the scenario should drive users to react to live events related to that scenario. This is true in sportive contexts and political debates, for example. We have made this clearer in the paper, specially in the Introduction. However, one good point of our work is that we do not need to know which events are taking place, we just observed the fact that in some moments some groups of users manifest themselves with more intensity than on other moments.

Also it is said that "the biases of the most vocal users are expected to be durable measures". Based on what? For how long? Is it scenario-dependent?

This is based on sociological studies that argue that humans tend to be biased and consistent in their opinions. Thus, we do not expect someone to change his supported football club, for example.

In the same direction, the authors state that "repeating a message indicates an endorsement of its content..". Is this based on what?

We have added a reference to a recent work [Guerra et al. 2011] that states that repeting the identical message in Twitter, through retweets, is a strong evidence of agreement with the original message. There are also other works that use endorsements for tasks such as content recommendation. We have added another reference for this.

Moreover the pre-algorithm task of classification of the bias of the users is delicate and deserves more attention of the authors. A heuristic based on the number of mentioned words is proposed. As a heuristic, it may contain imperfections. Many of the tweets, for example, are extremely contextual. Sometimes they did not mention the name of what they are talking about. For instance, it is possible to talk about a TV interview without mention the name of the interviewee. In one of the case studies, the authors use as heuristic the fact that "whenever the most cited team is three times higher than the sum of the mentions of all other team". Why three? How generic is this ad hoc criterion? What is the impact of the reliability of this assignment in the quality of the results? That is to say, how the algorithm reacts in noisy domains?

We have improved our pre-algorithm ground truth generation step by adding an extra criterion to generate ground truth: we now consider also the description that som users include in their Twitter profile which explicitly states their team preference. This has reduced the importance of the "3 to 1 citation rate" to one third of the users from the original submission, and we have found that 3-to-1 trade-off between false positives and false negatives. It is also important to stress that we also use tags that clearly indicates preference, such as #goJets and #voteDilma. We consider those tags a reliable source of ground truth, although we acknowledge this approach can bias a little bit our accuracy measures, as this users may be easier to analyze. Nevertheless, now we have in our dataset users who never mentioned his team or mention it in non-expected rations.

Who must attribute the bias in each scenario? How difficult this is? How it is possible to guarantee that for each set of messages (in a bin) there will exist users with known bias?

We do not need to strictly guarantee such constraint; when there are no users on an interval (bin), bias of users are not changed at all. In the case of peaks of activity, which is exactly when users react to live events, it is unlikely that no users with known bias will manifest.

Please define precisely what is "every component of a user bias" Page 6.
Component of user bias is the component of a vector.

What is labeled users in Table 1?
Labeled users are users whose class is known. We have changed, in Table I, "labeled" users for "known users".

Why the choice of 60 seconds of the interval? How to generalize this? Is there a significant impact if this interval is varied?
We have added a new Section (4.2) which shows the impact of varying this parameter in our results.

Please, precise how the recall and precision evaluation was computed.
We have added equations for Precision and Recall in Section 4.3.

2 Review 2

The main negative aspect is the proposal description (Section 3). The lack of details and examples makes it difficult to understand the proposal.
To make the description of our proposal clearer, we have added a new subsection (Sec. 3.4.3) which instantiates our algorithm with a running example. We also improved the description of each instruction of our algorithm in the paper.

Also, I am not convinced that the precisions achieved by the algorithm in the three scenarios are sufficient to state that it provided accurate results. The authors compare their precision results only with a random classifier.
We have changed our baseline comparison from a random classifier to a majority-class classifier. This guarantees that our approach is not generating falsely good results due to skewed class distributions.
We also would like to state that precisions, on average, are low because a high fraction of users manifest themselves very few times. However, we have shown that temporal information can provide useful information for users who manifest more frequently. We agree with the reviewer that we should not argue that our results are "highly accurate" and we changed the interpretation in text to state that we show that temporal information can be very useful if combined with other evidences, such as social ties.

In Section 3.3.1.: "A way to automatically define the known biases is to look at the content of a fixed subset of messages..". How this subset of messages is chosen?

It is chosen based on messages whose content is easy to be judged due the presence of tags.

In Section 3.3.2.: "It takes groups of messages one by one... For every group U, it processes the messages...". From this sentence, I understood that U is a group of messages. However, in Algorithm 1, U â GetUsersInAGroup(), i.e., U is a group of users. After all, what is U?

U is a multiset of users whose manifested themselves in an interval. We had added this explanation, with an example, in Section 3.3.

In Algorithm 1: What is the initial value of S?

S is initially the null vector (0,0...0,0). We have added this initialization step to Algorithm 1.

Why S â S + —U—.V?

To make our algorithm clearer, we changed —U— for k, and, for each interval, k is the number of users of known biases who posted in the interval. We multiply V for k because we want S to contain the sum of the vector of **all** users with known bias, so V must be added k times in each step.

Please, give more details about the algorithm implemented. Perhaps some examples would help.

We agree with the reviewer and to make the description of our proposal clearer, we have added a new subsection (Sec. 3.4.3) which instantiates our algorithm with a running example.

In Section 3.4.: "Thus, a bias in B is to be normalized so that the sum of the normalized user biases is the null vector". Why null vector?

Actually, this was a mistake in the text, inSection 3.4. A sentence stated the sum of normalized biases were the null vector, but actually it is the vector (1,1,...,1,1). The idea is to normalize user bias by the sum of bias of all known users, to correct class imbalacement.

In Section 4.1.: In fact, the data presented in Table 1 show that your proposal runs very quickly. Please, include informations about the computational resources used in the experiments conducted in this work.

We have added the information required by the reviewer, stating in Section 4.1 that we used a single core machine with 24GB RAM memory.

The algorithm proposed in this paper achieves precisions ranging from 45% to 75%. The authors concluded that the algorithm has provided accurate results

in all three scenarios, but, at first, a precision of 45% or 55% does not seem good to me.

Precisions, on average, are low because a high fraction of users manifest themselves very few times. However, we have shown that temporal information can provide useful information for users who manifest more frequently. We agree with the reviewer that we should not argue that our results are "highly accurate" and we changed the interpretation in text to state that we show that temporal information can be very useful if combined with other evidences, such as social ties.

Minor issue: Please, correct the date â2010-15-12â in Table 1.

Done.

3 Review 3

The authors compare the precision of their results only with random classification. However, comparison with competing approaches is relevant to show the accuracy reached by using temporal locality (instead of message contents). Also, computational time measures could reinforce the claim that the algorithm is scalable and feasible in real-time, very large streams.

We have changed our baseline comparison from a random classifier to a majority-class classifier. This guarantees that our approach is not generating falsely good results due to skewed class distributions.

- Section 3.3.1 â the authors affirm that, although message contents must be read in order to determine known biases, "it does not need to be performed on a huge quantity of messages". So, how many messages are necessary? In real-time applications, what is the impact of the number of messages and/or the period of time (e.g. do one need to get periods of high activity?)

- In Table I, what is the "avg time/interval"? Time to group the messages? Time to learn the group bias?

It is the whole time to process a group, which include grouping messages and learning bias from the group. We have added a sentence with this explanation in the text.

- Section 4.1 â Why is the user considered known if she/he cites a team three times more than the other teams? Why three times?

To improve that ad hoc criterion we have added an extra criterion to generate ground truth: we now consider also the description that some users include in their Twitter profile which explicitly states their team preference. It is also important to stress that we also use tags that clearly indicates preference, such as #goJets and #voteDilma. We consider those

tags a reliable source of ground truth, although we acknowledge this approach can bias a little bit our accuracy measures, as this users may be easier to analyze. Nevertheless, now we have in our dataset users who never mentioned his team or mention it in non-expected rations.

The explanation of the Algorithms is quite brief. Further details could make reading easier.

We have included further explanations of the steps of our algorithm, and have included a running example in Section 3.4.3. We think the running example improves a lot the understanding of our approach.

How does the proposed approach deal with simultaneous events? For instance: two ianeous soccer matches and a lot of unknown users commenting goals of one of the matches.

We added two sentences in Section 4.2 explaining how our algorithm deals with simultaneous events:

Note that simultaneous events can drive users from different viewpoints to manifest at the same time; those multiple events will help on discriminating those viewpoints from the viewpoints which are currently not related to those events. In general, however, most events should occur in isolation, for example, most of goals and touchdowns do not occur at the same time even if multiple matches are taking place simultaneously.

It seems there are several "parameters" that may impact the accuracy of the results, such as time intervals, heuristic to search in the microblogging platform and to define known biases, peaks of activity, "retweets", and so on. Further discussion on these issues are relevant to base the claim that temporal locality can be applied to a wide range of domains in different languages.

- Figures 4, 5 and 6: some curves can hardly be seen in printed version.

Finally, some references are incomplete.

We have completed two uncomplete references.

4 Review 4

The authors claim that the results are significantly different than a random choice, however this may not be good enough for a real application of the approach.

We agree with the reviewer; we have changed our comparison with a majority-class classifier and we added a comment in the Conclusions stating that our paper shows the usefullness of a new dimension, currently ignored in the literature – the temporal behavior of microblog users, and that this dimension can be combined with other evidences the literatura already

considers (such as social ties) to generate results which are better than when these characteristics are considered in isolation.

We have added a sentence in the Conclusions saying that, besides showing that it is possible to use temporal information to predict user inclinations, our work is useful to motivate researchers to use the temporal dimension to build and design applications that use it to customize and enhance user experience in social media. We think this is an important aspect of our work that was not highlighted in our original submission. For example, a possible application is to design an real-time recommendation algorithm that recommends content in real-time, or recommends users that tend to post at the same time.

The approach is trained (and evaluated through cross-validation) on users that have mentioned the name of a team or a candidate. Therefore, we know that there is temporal locality for those users. This says nothing about the users who never express their bias explicitly. How do I know if the system also finds their biases correctly?

He have improved our method of generating ground truth of users, which now considers the user description. So, we now evaluate users who never mention their teams explicitily in their tweets.

The system uses a very simple handcrafted classifier, which is obtained by binning the tweets into discrete time intervals. Why not use existing classification techniques that are well-known in the literature such as k-NN? (Finding the k-closest tweets using time as the distance function

This suggestion is interesting and we will leave it for future work. Our goal in this paper was to show that even a simple approach is able to capture meaningful patterns from temporal data in microblogs.

In Figure 1, I believe that the solid arrows are going in the wrong direction (if not, the figure does not make any sense to me).

We agree with the reviewer and changed the figure.

What are ACID principles? Areference should be given.

ACID (atomicity, consistency, isolation, durability) is a set of properties that guarantee database transactions are processed reliably. We have added a reference form ACID properties in the paper.

I do not like the justification for normalization that mentions the class imbalancement problem. I see this as a completely unrelated problem. I think the algorithm 1 would make more sense (and not require normalization) if they use some kind of formalism, like probabilities, instead of just counts. Another option would be to use a well-known classification method such as k-NN.

How are the average precision/recall obtained? In a micro or macro way?

We obtained precision and recall in a micro way. We have added equations for precision and recall in the paper.

I would like to see a comparison with the precision/recall obtained if we assign all tweets to the majority class. I expect the distribution of classes to be highly skewed in the sports domain so it could be that assigning to the majority class performs just as well as the temporal approach in these cases (and comparison against the random choice in this case would not be good enough).

This is a very good suggestion and we included those results in the paper, and our algorithm is still able to beat the majority-class classifier, specially in the sportive domains.

The approach is trained (and evaluated through cross-validation) on users that have mentioned the name of a team or a candidate. Therefore, we know that there is temporal locality for those users. This says nothing about the users who never express their bias explicitly. How do I know if the system also finds their biases correctly? It could be that these users have a different behaviour than the ones who express themselves explicitly and this would make the system useless (since for those users we already know the bias). One way of verifying this would be to ask the users who never expressed their bias explicitly if the inferred bias is correct.

Our new criterion of determining ground truth of user bias do not require anymore that users mention their team in tweets, so we now consider a more diverse set of users.

Best regards,

Pedro H. Calais Guerra, Loic Cerf, Adriano Veloso, Wagner Meira Jr. and Virgilio Almeida.