Identifying Automatic Posting Systems in Microblogs

Gustavo Laboreiro[1], Luís Sarmento[1,2], and Eugénio Oliveira[1]

[1] Faculdade de Engenharia da Universidade do Porto - DEI - LIACC, Portugal
[2] SAPO Labs Porto, Portugal
{gustavo.laboreiro,las,eco}@fe.up.pt

Abstract. In this paper we study the problem of identifying systems that automatically inject non-personal messages in micro-blogging message streams, thus potentially biasing results of certain information extraction procedures, such as opinion-mining and trend analysis. We also study several classes of features, namely features based on the time of posting, the client used to post, the presence of links, the user interaction and the writing style. This last class of features, that we introduce here for the first time, is proved to be a top performer, achieving accuracy near the 90%, on par with the best features previously used for this task.

Keywords: User-Generated Content, UGC, Microblogging, Twitter, Spam, Bot, Automatic Identification, User Classification, Stylistic Analysis, Noisy Text.

1 Introduction

Microblogging systems — of which Twitter is probably the best known example — have become a new and relevant medium for sharing spontaneous and personal information. Many studies and applications consider microblogs as a source of data, precisely because these characteristics can confer authenticity to results. For example, *trend detection* ([8]), *opinion-mining* ([6]) or *recommendation* ([4]).

Because of its popularity, Twitter is also part of the on-line communication strategy of many organizations, which use a Twitter account for providing updates on news, initiatives, commercial information (e.g. promotions, advertisements and spam) and various other types of information people may find interesting (like weather, traffic, TV programming guides or events).

Messages conveyed by these automatized accounts – which we will now refer to as *robot accounts* or, simply, *bots* – can easily become part of the stream of messages processed by information extraction applications. Since bots provide content aimed at being consumed by the masses instead of the personal messages that information extraction systems consider meaningful (for example, for trend detection), automatic messages may bias the results that some information extraction systems try to generate. For this reason, from the point of view of these systems, messages sent by bots can be considered noise.

L. Antunes and H.S. Pinto (Eds.): EPIA 2011, LNAI 7026, pp. 634–648, 2011.

The number of such robot accounts is extremely large and is constantly growing. Therefore, it is practically impossible to manually create and maintain a list of such accounts.

Even considering that the number of messages typically produced by a bot each day is not significantly larger than the number of messages written by an active user in the same period, we must remember that bots are capable of sustaining their publication frequency for longer periods than most humans (that can stop using the service temporarely or permanently after some time). Thus, in the long run, bots are capable of producing a larger set of messages than an active person.

In this work we propose a system that can identify these robot accounts using a classification approach based on a number of observable features related to activity patterns and message style. We evaluate its performance, and compare it with some of the more common approaches used for this task, such as the client used to post the messages and the regularity of new content.

1.1 Types of Users

Based on the work of Chu et al.[5], we start by distinguishing between three types of users. The term *human* is used to refer to users that author all or nearly all their messages. They usually interact with other users, post links on some of the messages, use abbreviations, emoticons and ocasionally misspell words. Many employ irregular writing styles. The subject of their messages can be different, but they tend to express personal opinions. Below we have examples of two human users:

- Who's idea was it to take shots of tequila? You are in so much trouble.
- I forgot to mention that I dropped said TV on my finger. ouchie.
- Heard that broseph. RT @ReggaeOCD: So bored with nothing to do. #IHateNotHavingFriends
- Just being a bum today. http://twitpic.com/4y5ftu

- aw, grantly:'destroys only happy moment in fat kids life' when talking about food :@
- @ryrae HAHAHAHAHHAHA :')
- JOSH IS IN SEASON 5 OF WATERLOO ROAD! WHEN DID THIS HAPPEN?

Bots, on the other hand, are in place to automate the propagation of information. The content is generally written by a person, although in some cases the entire process is automated (e.g. sensor readings).

We should note that what we are distinguishing here is more a matter of content than a matter of process or form. It is possible that an account where a person writes the message directly at the Twitter website be labeled as a bot, if the messages are written in the cold objective way seen in the examples below, from three different accounts.

- Social Security and Medicare to run short sooner than expected. http://on.cnn.com/lauSNv
- For Louisiana town, a collective gasp as it braces for floodwaters. http://on.cnn.com/mb481c
- Jindal: Morganza Spillway could open within the next 24 hours. http://on.cnn.com/j2jIBs

- 96kg-Bosak takes 7th place at the University Nationals
- 84kg-Lewnes takes 2nd to Wright of PSU and qualifies for the world team trials in Oklahoma Ciry
- Bosak loses his consolation match 0-1, 1-3 to Zac Thomusseit of Pitt.

- #Senate McConnell: Debt limit a 'great opportunity' http://bit.ly/kanlc9 #Politics
- #Senate Wisconsin Sen. Kohl to retire http://bit.ly/kQpAsS #Politics
- #Senate Ensign may face more legal problems http://bit.ly/mN7XsB #Politics

Many accounts are not run entirely by a person nor are they completely controlled by a machine. We label these mixed accounts as *cyborgs*, the term used by Chu et al. [5], that describes them as a "bot-assisted human or human-assisted bot". For example, an enterprise can have an automatic posting service, and periodically a person provides the user interaction to maintain a warmer relation with the followers, and foster a sense of community. Another possibility occurs when a person uses links to websites that post a message on the account of that person (for example, "share this" links). If these pre-written content are noticeable among the original messages, the user is labeled as a cyborg. If barely no original content is present in the user's timeline, they will be considered a bot. Below we show examples of a cyborg account.

- Explore The Space Shuttle Era http://go.nasa.gov/gzxst5 and immerse yourself in the Space Shuttle Experience http://go.nasa.gov/iHVfGN
- Track the space shuttle during launch and landing in Google Earth using real-time data from Mission Control http://go.nasa.gov/mwO9Ur
- RT @Rep_Giffords: Gabrielle landed safely @NASAKennedy. For more details go to www.fb.me/GabrielleGiffords. #NASATweetUp
- Space shuttle Endeavour's preferred launch time moved two seconds later! Now 8:56:28 a.m. EDT Monday.
- @Angel_head NASA frequently tweaks the shuttle launch time by seconds based on the latest space station tracking data to use the least fuel.

As we will explain next, we used these guidelines to construct a Ground Truth that will be used in our experiments. The details of this task are given in Section 4.1.

To address the problem of automatic posting, we study different sets of features that allow us to classify Twitter users into the three user categories described. These features explore characteristics exhibited by the users, such as

their posting times, the microblog client application they use, their interaction with other users, the content of their messages, and their writing style. The main goal of the work presented in this paper is to access the usefulness and robustness of the different types of features proposed. We discuss the features in Section 3.

We describe our experiment and its parameters in Section 4 and evaluate our results in Section 5. In Section 6 we present our conclusions and future work.

2 Related Work

Most literature addressing the identification of automated systems in microblogs is related with the detection of spam. While there is some overlapping between spam and automated posting systems (spammers often employ automation to help them in their work), we feel that the problem we are addressing is much more general.

Wang [10] presents an effort to detect spamming bots in Twitter using a classification based approach. The author explored two sets of features: (i) information about the number of followers, friends and the follower per friend ratio for the social network aspect of Twitter; and (ii) information about duplicate content and number of links present in the last 20 messages of a given user account. The author used a manually annotated corpus to train a Naive-Bayes classifier. The classifier achieved slightly over 90% Precision, Recall and F-measure in a 10-fold cross validation experiment. However, since the training corpus was biased towards non-spam users (97% of the examples), any classifier that only reported "non-spam" would be almost always correct, so results are not really significant.

Grier et al. [7] analyze several features that indicate spamming on Twitter. They looked for automated behavior by inspecting the precision of timing events (minute-of-the-hour and second-of-the-minute), and the repetition of text in the messages across a user's history. They also studied the Twitter client application used to write the messages, since some allow to pre-schedule tweets at specific intervals.

Zhang and Paxson [11] present a study where they try to identify bots by looking only at the minute-of-the-hour and the second-of-the-minute. If the posting times are either too uniform or not uniform enough, there is the possibility of the account being automated. This analysis is similar to the one present in Grier et al. [7], a work where Zhang and Paxson participated.

The authors present no validation of their results (since it is not possible to determine for sure if the account is automated or not). However, they claim that "11% of accounts that appear to publish exclusively through the browser are in fact automated accounts that spoof the source of the updates".

Chu et al. [5] propose to distinguish between humans, bots or cyborgs, but much of the effort was put into spam detection. They claim that more sophisticated bots unfollow users that do not follow back, in an effort to keep their friends to followers ratio close to 1, thus reducing the effectiveness of features based on the social network of Twitter. However, by discarding the uncertain

and ambiguous examples from their Ground Truth set, the authors seriously reduced the usefulness of their results.

Contrary to previous work, we focus on a problem that is much more generic than spam-detection, since a very large number of bots belong to newspapers and other organization, which are voluntarily followed by users. Our goal is to separate potentially opinionated and highly personal content from content injected in the Twittosphere by media organizations (mostly informational or promotional). One other point where we distinguish ourselves from the mentioned works is in our attempt to expand the set of features used in the detection, now including a vast array of stylistic markers. In our opinion, this opens a new field for exploration and study.

3 Methodology

Most of our discussion is centered around distinguishing human users and bots. We propose five sets of features, described below, that are intended to help to discriminate between these two poles. A cyborg user, by definition, exhibits characteristics typical of both classes of users.

3.1 Chronological Features

One of the characteristics of automatic message posting systems is that they can be left running indefinitely. Therefore, we can expect to see different chronological patterns between human users and bots. To address these points, we defined a number of features, divided into the four following sub-classes.

Resting and Active Periods. Constant activity throughout the day is an indication that the posting process is automated or that more than one person is using the same account — something we expect to be unusual for individual users. Figure 1 was drawn using information from our manually classified users (described in Section 4.1). It shows how human and cyborg activity is reduced between 1 and 9 AM. Bot activity is also reduced between 11 PM and 7 AM, but it never approaches zero.

At the same time, other things can keep people from blogging. This can lead to a certain hour of preferred activity, such as evenings, as suggested in Figure 1. Bots appear to have a more evenly spaced distribution across the day, with smaller fluctuations in the level of activity. This can be a conscious choice, to allow more time for their followers to read each post.

Since we tried to limit our crawling efforts to Portugal, most of the observations are expected to fall within the same time zone (with the exception of the Azores islands, which accounts for 2% of the population). We believe that the problem of users in different time zones cannot be avoided completely. For example, we do not expect users to correct their Twitter profile when traveling.

To detect the times at which the user is more or less active, we define 24 features that measure the fraction of messages they posted at each hour. These values should reflect the distributions represented at Figure 1. We also analyze

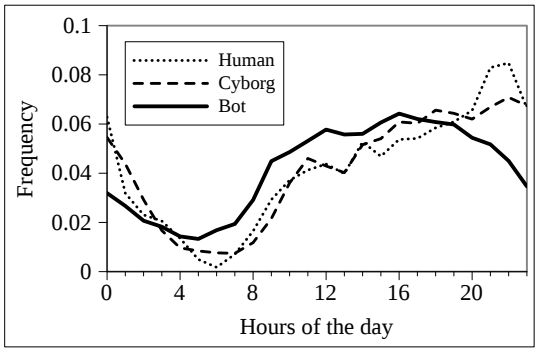

Fig. 1. Comparison of Twitter activity between bots, cyborgs and humans as a function of the hour

the average and standard deviation of these values. We expect that the standard deviation of a bot is lower than that of a human.

Finally, we register the 10 hours with the highest and lowest activity, and the average number of messages that the user posts per day (as a floating-point number).

Long term Activity. Days are not all equal. This is true for both humans and bots. For example, as shown in Figure 2, most activity happens at Thursdays and Fridays. This trend matches the result published by Hubspot [1].

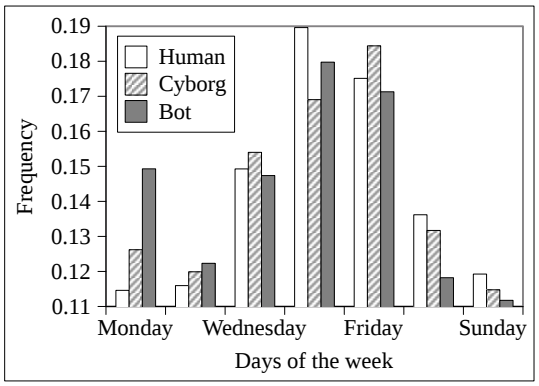

Fig. 2. Comparison of Twitter activity between bots, cyborgs and humans as a function of the day of the week

We can see that bots are less active during the weekends, while they dominate on Mondays and lead on Tuesdays.

We define seven features related to the frequency of the messages posted across each day of the week, that should reflect the proportions in Figure 2. We also calculate the workday and weekend posting frequency, and rank the days of the week by the frequency of posting.

Inactivity Periods. From direct observation of Twitter messages, we can see that bots tend to be more regular on their updates than humans. It is known that irregular accounts can lose popularity quickly. At the same time, normal people need to rest, get occupied with other matters, and can lose interest in blogging for some time.

To make use of this information, we measure the periods of inactivity in minutes, and record the length of the 10 longest intervals, in decreasing order. We also calculate the average and standard deviation of all these values. From our observations, we expect that bots will have lower variation in inactivity periods (lower standard deviation) and a higher average.

For example, considering a user that only blogged at 1 PM, 2 PM, 3:30 PM and 7 PM on the same day, we would have the following features:

Feature name	Value (minutes)
top_inactive_period_1	210.00
top_inactive_period_2	90.00
top_inactive_period_3	60.00
average_top_inactive_period	120.00
standard_deviation_top_ inactive_period	79.37

Humans are unable to match the speed at which bots can create new messages. For this reason, we also calculate the analogous features for the minimum inactivity periods (i.e. the 10 shortest inactive periods).

Posting Precision. Since some automatic posting systems work based on a fixed periodicity (e.g. TV programming guides), we decided to calculate the frequency of messages that are created at each minute (60 features) and second (another 60 features). This approach is a simpler version of other works [7,11].

For a human, we expect their posts to be evenly spread across both these measurements. Some bots, on the other hand, are expected to concentrate their activity around the 0 seconds mark. They may also do the same around some particular minutes (e.g. 0 or 30).

As before, we also calculate the average and standard deviation of these measurements, where humans should result in a lower average and higher standard deviation compared to bots.

For both minutes and seconds we take note of the 10 most frequent values — that is, when most activity occurs.

3.2 The Client Application Used

It makes sense that the Twitter client used to post the messages be a relevant aspect in identifying automated processes. There are many clients and methods

of accessing the microblogging system (e.g. web interface, several applications, etc.). From the point of view of automation, some of these methods are easier or more convenient than others. Also, most microblogging systems have an open API, meaning that it is possible to interact with them directly. In Twitter, unregistered clients are identified as "API", while those that were registered are identified by their name.

We track the number of different clients used to post the messages, and the proportion of messages posted with each client. Cyborgs are expected to have the largest variety of clients used (as they usually post automatically from several sources). Some humans can use more than one client, for example, a mobile client and the website. Bots, on the other hand, can adhere to a single, exclusive client that is tied with their on-line presence; or may use a general client that imports messages from an RSS feed, for example.

3.3 The Presence of URLs

We can make a distinction between two types of bots: information bots, which only intend to make their readers aware of something (e.g. weather forecast, TV scheduling and traffic information), and link bots, whose main purpose is to generate traffic towards their website (e.g. news, advertisements and spam). Information bots usually don't have URLs in their messages, while some link bots are capable of truncating the text of the message to make room for the URL. Most URLs shared by a bot usually have the same domain, i.e. they were all created by the same URL shortening system, or point to the same website.

Humans are also capable of introducing many URLs, but our observation reveals that cyborgs are more likely to do so; and to vary the domains of said URLs. We can observe both types of linking behaviour represented in the bot and human examples presented in the introduction, in Section 1.1.

We defined a feature that represents the ratio of URLs shared per total of messages written, and also keep track of the proportion of the domains associated to the URLs.

3.4 User Interaction

Bots usually have one main objective that is to spread information regularly. While they may be programmed to do more complex actions (such as follow other users), automating user interaction can have undesired repercussions for the reputation of the account holder. Thus, *reblogs*[1] (to post a copy of another user's message) and *replies* (directing a message at a user) are shuned by most bots. The main exception are some spamming bots, that send several messages directed at users [7]. To include the name of other users in the message (*mentions*) also seems to be uncommon in automated accounts.

However, a number of users also avoid some types of interaction, such as the ones previously mentioned. Therefore, while we expect this information to help identify humans, they may be less helpful in identifying bots.

[1] Called "retweets" on Twitter, often shortened to "RT".

With our features we keep track of the proportion of reblogs, replies and mentions, as well as the average number of users mentioned per message written.

3.5 Writing Style

Stylistic information has been successfully used to distinguish the writing style of different people on Twitter [9]. Thus, we believe it to be helpful in distinguishing between automated and non automated messages since, as observed in the examples in Section 1.1, these users adopt different postures. The austere writing style may help with the readability, and also credibility, associated to the account, while many humans do not seem too concerned about that.

We identify the frequency (per message) of a large number of tokens, as listed below.

Emotion Tokens. Bot operators wish to maintain a serious and credible image, and for this reason avoid writing in a style too informal (or even informal). We collect information on the use of various popular variations of smileys and "LOLs".

We also try to identify interjections. While this part of speech is culturally dependent, we try to identify word tokens that have few different letters compared with the word length — if the word is longer than 4 characters, and the number of different letters is less than half the word length, we consider it an interjection.

These three stylistic features were the most relevant features mentioned by Sousa-Silva et al. [9]. Below, we can see example messages containing many emotion tokens:

- we talked before..... on twitter. **HAHAHAHAHA** RT @Farrahri: @Marcology **LOL** she smiled at me! **Hehehe**, jealuzzz not?
- **riiiiiight**.... im **offffffffffffff**!!!!! bye bye
- RT **yessssssss**! That is my **soooong**!!!! @nomsed: You got the **looove** that I **waaaant** RT @LissaSoul: U got that BUTTA **LOOOOooooVVVEEE**!

Emotion can also be expressed through punctuation, but we include those features in the punctuation feature group, below.

Punctuation Tokens. Humans vary widely in regards to their use of punctuation. Many are not consistent across their publications. This is in opposition to bots, that can be very consistent in this regard.

Punctuation can often be used as a separator between the "topic" and "content" of the message, as can be seen on the first example on Table 1 . Different sources of information may structure their messages differently. Therefore one bot may use more than one separator.

We also notice that some bots publish only the headline of the article they are linking to. These articles are usually blog posts or news at a news website. Since headlines usually do not include a full stop, this feature receives a very low frequency (as seen in example 1).

Table 1. Examples of bot Twitter messages making use of punctuation for structural purposes

Example	Text
1	Tours**:** Brian Wilson should retire next year http://dstk.me/Oi6
2	Gilberto Jordan at Sustain Worldwide Conference 2011**:** Gilberto Jordan, CEO of Grupo André Jordan, is the only spea**...**http://bit.ly/mOOxt1

Question or exclamation marks are usually infrequent in news bots, or bots looking for credibility [2]. Some bots truncate the message to make room for the URL, signaling the location of the cut with ellipsis (some using only two dots). We can see an example on the second example on Table 1.

We measure the frequency of occurrences of:

- Exclamation marks (single and multiple);
- Question marks (single and multiple);
- Mixed exclamation and question marks (e.g. "!?!?!?!?!!?!?!!");
- Ellipsis (normal [i.e. "..."], or not normal [i.e. ".." or "...."]);
- Other punctuation signs (e.g. full stop, comma, colon, ...);
- Quotation marks;
- Parenthesis and brackets (opening and closing);
- Symbols (tokens without letters and digits); and
- Punctuation at the end of the message (both including and excluding URLs).

Word Tokens. This group of tokens is kept small for the sake of language independence. We begin by tracking the average length of the words used by the author. We also define features that track the frequency of words made only of consonants (that we assume to be abbreviations most of the time), and complex words. We consider complex words as those having more than 5 letters and with few repeated characters (more than half). Thus "current" (7 characters in length, 6 different characters) is a complex word, while "lololol" (7 characters in length, 2 different characters) or "Mississippi (11 characters in length, 4 different characters) do not fit the definition.

Word Casing. Bots are usually careful in the casing they use. Careful writing aids with the image that is passed through. We measure the frequencies with which the following is used:

- Upper and lower cased words;
- Short (≤ 3 letters), medium (4–5 letters) and long (≥ 6 letters) upper cased words;
- Capitalized words; and
- Messages that start with an upper cased letter.

Quantification Tokens. We track the use of some numeric tokens. Dates and times are comonly used to mention events. Percentages can be more common on news or advertisements.

- Date (e.g. "2010-12-31" or "22/04/98");
- Time (e.g. "04:23");
- Numbers;
- Percentages; and
- Monetary quantities (e.g. "23,50€", "$10" or "£5.00").

Beginning and Ending of Messages. Some accounts post many messages (some times all or near all their messages) using one or a small number of similar formats. This behavior is specific of bots, that automate their posting procedure. Below we can see two examples.

- **#football** Kenny Dalglish says Liverpool will continue conducting their transfer business in the appropriate manner. http://bit.ly/laZYsO
- **#football** Borja Valero has left West Brom and joined Villarreal on a permanent basis for an undisclosed fee. http://bit.ly/iyQcXs

- **New post:** Google in talks to buy Hulu: report http://zd.net/kzcXFt
- **New post:** Federal, state wiretap requests up 34% http://zd.net/jFzKHz

To determine the pattern associated with the posts, we calculate the frequency with which messages begin with the same sequence of tokens (excluding URLs and user references, that frequently change between messages). We define tokens as words, numbers, punctuation signs, emoticons and other groups of symbols that have a specific meaning.

We group all the messages by their first token. For each group with two or more messages, we store their relative proportion in a feature related to the token. We also register the 10 highest proportions found, in descending order. This entire process is then repeated, looking at the first two tokens, then the first three, and so on.

Once complete, we take note of the largest number of tokens seen, and repeat the entire process, looking at the endings of the messages.

This procedure results in a number of features that are very specific. In the case of humans we collect a relatively small number of features, as their messages can be varied. Some bots will reveal a pattern that is used for *all* their messages (e.g. see the last bot examples in Section 1.1). In the case of cyborgs, it is very useful to detect a number of patterns such as "I liked a @YouTube video *[URL]*" or "New Blog Post ...".

Below we can see examples of messages where this approach is useful. The first two messages are from a bot account, while the last two were taken from a cyborg account.

4 Experimental Set-up

Our aim is to compare the level of performance provided by the five sets of features described in Section 3. First we create a Ground Truth by classifying a number of Twitter users manually. This data is then used to both train and test our classification system.

4.1 Creation of a Ground Truth

In late April 2011 we started a Twitter crawl for users in Portugal. We considered only users who would specifically state that they were in Portugal, or, not mentioning a known location, that we detect to be writing in European Portuguese. This collection started with 2,000 manually verified seeds, and grew mostly by following users that are referred in the messages. In this way our collection moved towards the more active users in the country. However, there is no guarantee that we have been collecting all the messages from any of the users.

At the moment we have over 72 thousand users and more than 3 million messages. From this set, we selected 538 accounts that had posted at least 100 messages, and 7 people were asked to classify each user as either a human, a bot or a cyborg, in accordance with our guidelines, as described in Section 1.1.

The annotators were presented with a series of user accounts, displaying the handle, a link to the Twitter timeline, and a sample of messages.

Since the users presented to the annotators were randomly selected, not every annotator saw the users in the same order, and the sample of messages for each user was also different. For each user, we considered the classification that the annotators most often attributed them. In the case of a tie, we asked the annotator to solve them before ending the voting process.

To finalize the voting process, we asked one eighth annotator to solve the ties between annotators. In the end we were left with 2,721 votes, 95% of which from the 4 main annotators, that we used to calculate the agreement. Looking only at the 197 users that were classified by all 4 main annotators, we get a substantial 0.670 Fleiss' kappa value, showing adequate reliability in the classification.

Figure 3 shows the distribution of the users across all three categories. We can see that humans dominate our collection of Twitter (448), while cyborgs were the least numerous (22). In total we identified 68 bots.

4.2 The Classification Experiment

We randomly selected our example users from our manually classified examples. To have a balanced set, we limited ourselves to only 22 users of each type, randomly selected before the experiment. To handle the automatic classification, we opted for an SVM due to its ability to handle a large number of features. We opted for the libSVM [3] implementation.

For each user we selected up to 200 messages to analyze and create the features. Due to the chronological features (Section 3.1), we selected only sequential messages in our collection.

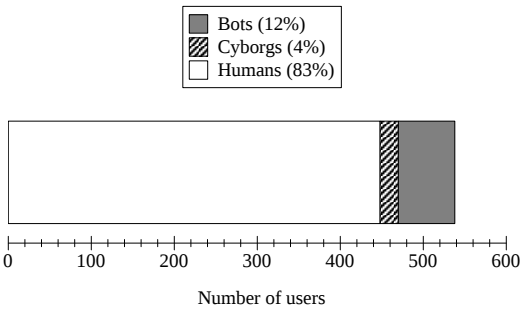

Fig. 3. Distribution of the 538 users in the three classes

We used the radial basis function kernel from libSVM, allowing it to look for the parameters that best fit the data, and normalized the values of the features, allowing for more accurate results. We measured the results using the accuracy, i.e. the ratio between correct classifications and total classifications.

We opted for a 2-fold cross validation system, where we select 11 users of each type to be used in the training set, and the other 11 were part of the testing set. This allows enough testing messages to provide adequate granularity in the performance measurement, and a more reasonable number of messages to train the SVMs. We repeat each experiment 50 times (drawing different combinations in the training and testing set).

Given that we are using a balanced set of examples, we expect that a random classifier would be correct 1/3 of the times. We will be considering this as the baseline in our analysis.

5 Results and Analysis

Our results are represented in Figure 4, representing the minimum, lower quartile, median, upper quartile and maximum accuracy across the 50 runs.

We separate the results in two groups: the first group, that never reaches 100% accuracy, and the second group, that does.

In the first group, the user interaction features outperformed the chronological features, that had two poor runs. However, none of them shows performance similar to the other feature sets.

In the second group, the stylistic features presents the best results, with median accuracy 97%. The feature that identifies the client application also performs adequately, but twice failed 7 or 8 of the 33 examples. The URL features showed more stable results than the client information feature, but generally failed in more cases. Finally, using all the features combined yielded very good results, with 97% median (and 97% upper quartile, hence overlapping in Figure 4), failing once in 5 of the examples.

Over 26,000 features were generated during the experiments, most of them encode stylistic information. While in a large group they can be quite powerful

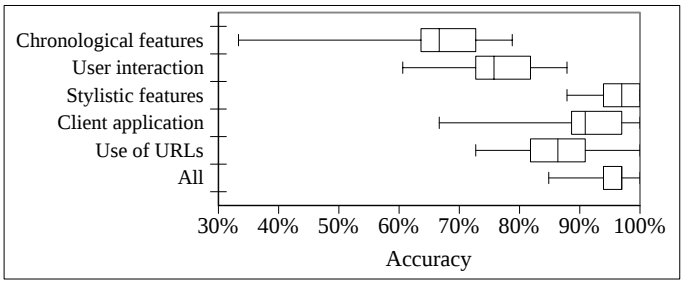

Fig. 4. Box plot showing the results for the classification of users, using 50 2-fold cross validation runs. The limits of the boxes indicate the lower and higher quartile. The line inside the box indicates the median. The extremities of the lines represent the minimum and maximum values obtained.

(as shown), each of these features carries little information. This is in contrast with the URL, user interaction and client application features, where a small number of features can contain very meaningful information.

Most features related with the client application, work almost like a database of microblog applications. That is, except for the number of different clients used, we are only recalling the identification strings present in the training messages. In the presence of an unknown client, the classifier has little information to work with. Hence the cases with low accuracy.

The features related with the URLs and with user interaction obtain information from the presence or absence of certain elements. However, in our implementation we could not encode enough information to address all the relevant cases, especially in the case of user interaction.

It is unfortunate that we are unable to compare our results with other approaches, mentioned in Section 2. There are three reasons for this: (i) their work has a different goal (i.e. spam detection); or (ii) the authors do not provide a quantification that we could use for comparison; or (iii) we consider that their experiment is biased (e.g. excluding some messages because they are more difficult to classify).

6 Conclusion and Future Work

We have shown that automatic user classification into either human, cyborg or bot — as we have defined them — is possible using standard classification techniques. In particular, as we have supposed, stylistic features can be a reliable indicator in this type of classification. In fact, they achieved results as good or better than other, more frequently used, indicators of automatic activity.

Basing the user classification in the client application used raises two problems: first, some applications can have mixed using (as Chu et al. [5] and Grier et al. [7] point out); and second, dealing with the large amount of different clients

is difficult. For example, we counted 2,330 different clients in our 73,848 users database (around 1 different client for every 32 users). Thus, while fast and simple, this approach does not appear to hold on its own, and should be combined with other approach.

In the future, we would like to improve our chronological features by adopting the same method Zhang and Paxson used [11,7], as our minute-of-the-hour and second-of-the-minute approach was, perhaps, too simplistic. We would also like to study the scalability of the stylistic approach, as they generate a large number of new features.

References

1. Burnes, R.: When do most people tweet? at the end of the week (January 2010), http://blog.hubspot.com/blog/tabid/6307/bid/5500/When-Do-Most-People-Tweet-At-the-End-of-the-Week.aspx
2. Castillo, C., Mendoza, M., Poblete, B.: Information credibility on twitter. In: Proceedings of the 20th International Conference on World Wide Web, WWW 2011, pp. 675–684. ACM, New York (2011)
3. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines (2001) software, http://www.csie.ntu.edu.tw/~cjlin/libsvm
4. Chen, J., Nairn, R., Nelson, L., Bernstein, M., Chi, E.: Short and tweet: Experiments on recommending content from information streams. In: ACM Conference on Human Factors in Computing. Association for Computing Machinery, Atlanta, GA (04/10/2010)
5. Chu, Z., Gianvecchio, S., Wang, H., Jajodia, S.: Who is tweeting on twitter: human, bot, or cyborg? In: Gates, C., Franz, M., McDermott, J.P. (eds.) Proceedings of the 26th Annual Computer Security Applications Conference, ACSAC 2010, pp. 21–30. ACM, New York (2010)
6. Dey, L., Haque, S.M.: Opinion mining from noisy text data. International Journal on Document Analysis and Recognition 12, 205–226 (2009)
7. Grier, C., Thomas, K., Paxson, V., Zhang, M.: @spam: the underground on 140 characters or less. In: Proceedings of the 17th ACM Conference on Computer and Communications Security, CCS 2010, pp. 27–37. ACM, New York (2010)
8. Mathioudakis, M., Koudas, N.: Twittermonitor: trend detection over the twitter stream. In: Proceedings of the 2010 International Conference on Management of Data, SIGMOD 2010, pp. 1155–1158. ACM, New York (2010)
9. Sousa-Silva, R., Laboreiro, G., Sarmento, L., Grant, T., Oliveira, E., Maia, B.: Twazn me!!!; automatic authorship analysis of micro-blogging messages. In: Procedings of the 16th International Conference on Applications of Natural Language to Information Systems (June 2011)
10. Wang, A.: Detecting spam bots in online social networking sites: A machine learning approach. In: Foresti, S., Jajodia, S. (eds.) Data and Applications Security and Privacy XXIV. LNCS, vol. 6166, pp. 335–342. Springer, Heidelberg (2010)
11. Zhang, C.M., Paxson, V.: Detecting and Analyzing Automated Activity on Twitter. In: Spring, N., Riley, G.F. (eds.) PAM 2011. LNCS, vol. 6579, pp. 102–111. Springer, Heidelberg (2011)