Computers in Industry 78 (2016) 3–15

Contents lists available at ScienceDirect

Computers in Industry

journal homepage: www.elsevier.com/locate/compind

Analysing and evaluating the task of automatic tweet generation: Knowledge to business

Elena Lloret *, Manuel Palomar

University of Alicante, Department of Computing Systems, Apdo. de correo 99, E-03080 Alicante, Spain

ARTICLE INFO

Article history:
Received 3 December 2014
Received in revised form 11 October 2015
Accepted 24 October 2015
Available online 7 November 2015

Keywords:
Natural language processing
Text summarisation
Natural language tweet generation
User study
Linguistic analysis
Descriptive statistics

ABSTRACT

In this paper a study concerning the evaluation and analysis of natural language tweets is presented. Based on our experience in text summarisation, we carry out a deep analysis on user's perception through the evaluation of tweets manual and automatically generated from news. Specifically, we consider two key issues of a tweet: its informativeness and its interestingness. Therefore, we analyse: (1) do users equally perceive manual and automatic tweets?; (2) what linguistic features a good tweet may have to be interesting, as well as informative? The main challenge of this proposal is the analysis of tweets to help companies in their positioning and reputation on the Web. Our results show that: (1) automatically informative and interesting natural language tweets can be generated as a result of summarisation approaches; and (2) we can characterise good and bad tweets based on specific linguistic features not present in other types of tweets.

1. Introduction, context and motivation

In the current digital knowledge society, the overload of information has become a problem to companies, which cannot cope with all the available information. As a consequence, companies may not be exploiting the Web, and taking advantage of it accordingly, thus affecting key aspects, such as their visibility, reputation, marketing campaigns, customer's feedback, etc. With the birth of the Web 2.0, there has been a shift in the way the information is produced and consumed by users and companies. The Web 2.0 has established a wide range of on-line mechanisms and platforms through which companies can obtain direct feedback from users. These mechanisms (e.g., reviews, social networks) allow users to freely express their comments about companies and the products/services they offer, thus requiring the effective management of a large number of adapted contents, formats, and interaction patterns [1]. Companies have envisaged the great potentiality of the communication through the Web 2.0 and even there have been attempts to integrate these channels into ERP platforms [2]. Moreover, companies have created their own social network profiles, e.g., in Facebook or Twitter, in order to increase their visibility, and maximise their interaction with customers.

With more than 241 million active users per month,[1] 184 million of which uses Twitter through their mobile device, and more than 500 million tweets daily,[2] Twitter[3] has become an excellent social media for on-line real-time news attention.[4] The length restriction imposed on tweets (140 characters) force messages to be concise, though it is also possible to link out to external information to enrich the tweet. Moreover, hashtags (e.g., #UA_Universidad) allow to categorise information, to identify the trending topics, and more importantly to enable a rapid on-line information flow. According to [3] one of the key success factors of Twitter is that it is an appropriate channel to communicate in short messages and share information regardless of time and place. Moreover, Twitter has become a means of electronic Word of Mouth communication (eWOM) [4], where one of its main usages is information distribution [5], that is spread very quickly reaching a high number of users in real time.

Companies are concerned about what their customers think about them, and in this manner, it is really important for them,

* Corresponding author.
E-mail addresses: elloret@dlsi.ua.es (E. Lloret), mpalomar@dlsi.ua.es (M. Palomar).

http://dx.doi.org/10.1016/j.compind.2015.10.010
0166-3615/

[1] http://expandedramblings.com/index.php/march-2013-by-the-numbers-a-few-amazing-twitter-stats/#.U0zftVfjDYM [last access June 2015].
[2] http://www.telegraph.co.uk/technology/twitter/9945505/Twitter-in-numbers.html [last access June 2015].
[3] http://twitter.com/.
[4] http://pando.com/2014/02/06/facebook-vs-twitter-who-wins-the-battle-for-our-social-attention/ [last access June 2015].

what and how information is delivered on the Web, since this may have a direct influence on their popularity and branding, affecting their positioning and reputation, or attracting/discouraging new potential customers. In the context of Twitter, the information to be expressed should be to the point, very clear and concise. This will benefit the impact on their business strategy, and will improve the relationship with customers, thus being able to personalise the information, as well as to improve marketing campaigns.

The level of maturity reached by state-of-the-art Natural Language Processing (NLP) techniques can support companies in delivering, managing and analysing on-line textual information. Current NLP applications, such as information retrieval, sentiment analysis or text summarisation could help companies to monitor relevant information about them, classify it, and obtain the key ideas. Specifically, when it comes to information delivery, text summarisation techniques could be used for automatically generating candidate micro- or ultra-concise summaries in the form of a tweet [6,7]. This task would be similar to headline generation [8,9], but in the current context of the Web 2.0, and in particular applied to Twitter, thus obtaining natural language tweets.

In the process towards the automatic generation of natural language tweets, a crucial stage is to know how users perceive them, and whether there are any linguistic features leading to the best and worst generated tweets. This will allow companies to be aware of the suitable language that would help to catch users' attention without negatively affecting its informativeness. Furthermore, the analysis of both issues would benefit communication strategies for companies, who need to be strategic in designing and executing their tweets [10].

Therefore, the main objective of this paper is to conduct a deep study on user's perception of tweets through the analysis of approximately 1600 tweets generated either by humans (i.e., manually), or by seven current text summarisers (i.e., automatically). Our study will be focused on analysing two key issues of a tweet: its informativeness and interestingness, as well as determining the set of linguistic features that contribute to produce good and not so good tweets. In particular, the research questions to answer are: (1) do users equally perceive tweets that have been manually generated in comparison to the automatic ones?; and (2) what linguistic features should or should not a good tweet have in order to be interesting, as well as informative for the user? Both, the identification of interesting and useful contents from large text-streams is a crucial issue in social media [11], and they have been widely employed for evaluation purposes in the context of Twitter [12–14]. Whereas for the first question, we use descriptive statistics for analysing in detail the assessment provided by different users, in the second question, we will collect a sample of several types of tweets and analyse in-depth their linguistic features, and main differences.

Moreover, our research work will be carried out from a multilingual perspective (for English and Spanish) with the purpose of determining if the language and the manner in which the tweet was generated have any influence on the user's perceptions. This intermediate research is framed within the overall research of automatically extracting and generating natural language tweets from external news documents talking about a company, product, etc. in order to help companies improve their positioning and reputation on the Web.

The results obtained from this research show that: (1) state-of-the-art summarisers are capable of generating good natural language tweets, that are informative as well as interesting, and that could be an alternative to manual generated tweets; and (2) it is possible to distinguish and characterise good and bad tweets based on different linguistic features that are not present in other types of tweets.

The remainder of this paper is organised as follows. Section 2 presents a literature review of the relevant research work related to the topic of this article. Section 3 explains the research methodology and questions that we want to analyse within the scope of this paper. It also provides information about the initial dataset that is used for conducting all the analysis, together with the NLP and statistical tools employed. Sections 4 and 5 show and discusses the findings and results of the analysis with regard to each of the proposed research questions. Finally, Section 6 presents the conclusions and final considerations, as well as several suggestions for future investigation.

2. Related work

Recently, Twitter has become a valuable source of data for research in NLP. The vast amount of data that each day millions of users and companies exchange through this platform has made it possible the analysis and processing of this textual genre, thus becoming necessary to analyse and exploit suitable techniques to filter out/discard irrelevant information, as well as to design effective and appealing communicative streams.

Natural language generation and text summarisation can help to achieve such challenging goals. The current difficulty associated to building natural language generation systems [15] and our considerable experience in text summarisation for extracting key ideas [16–18] has led us to address this study from a summarisation perspective rather than from natural language generation, even though generating natural language and applying it to Social Media (e.g., Twitter) would be our ultimate long-term goal.

In the literature, text summarisation techniques have been employed in the context of Twitter mainly for summarising tweet streams related to the same topic or event. Some examples of this type of approaches can be found in [19–23]. Different techniques such as phrase reinforcing algorithm or TF-IDF are employed, among others. Of all these approaches, we would like to highlight on the one hand, the approach proposed in [20], since it includes the novelty of taking into account not only the tweets themselves, but also the information linked by such tweets, and a combination of both of them. Whereas in most of the research works, tweets which come from user-generated content, are treated as they are, here, the authors apply a normalisation process to transform them into standard English. This is an important stage, because traditional NLP tools may fail when no standard language is provided [24]. Concerning the summarisation stage, the authors employed a concept-based optimisation approach for selecting informative sentences while minimising the redundancy. In this approach, the relevant sentences were determined based on the maximum number of concepts covered. The results indicate that the combination of normalised tweets and Web content was the best performing approach, beating the results obtained in [21]. On the other hand, [23] proposes an interesting novel aspect for Twitter event summarisation, which takes into account subjective information for generating a summary from different perspectives users may have on the same event. The authors focused on sport events, so they considered the fans' viewpoints for their approach. Given a set of tweets related to a sports event, the first step was to extract the ones referring to the teams involved in the event, and classify and group them with respect to the team it was supporting. Then, a topic detection algorithm was employed for returning up to ten topics for each event being considered, and they were compared with the comments related to the same event but belonging to an external information source (e.g., BBC comments). To select the closest topic, the cosine similarity measure was employed. Finally, for the selected topic, a small set of representative tweets of each of the groups was extracted.

Despite the number of research works producing summaries from Twitter data, there are only a few aiming at producing the opposite: a tweet as a summary of a heap of information. This task

could be considered similar to the traditional headline generation task; the aim of which is to summarise the key information in a single-sentence. Along different editions of DUC competitions[5] there was a specific task aiming at producing headlines no longer than 50 words. The techniques employed for producing these very short summaries included the use of lexical and named entities chains [25], information about the topic [26], parsing and trimming [27] or language models [28]. Recent research has focused on applying the headline generation task to produce titles [29], image captions [30], or even story highlights [31].

More recently, a similar idea was applied using user-generated content, where text summarisation techniques were used to generate micro- or ultra-concise summaries automatically [6]. In this approach, opinionated information was taken into account, and a tweet was a summary of a key opinion in a set of reviews. For identifying key opinions in the text, the techniques employed were based on Web Ngrams, obtaining good results when evaluating the automatic ultra-concise summary through a readability assessment.

Furthermore, the importance of generating interesting and catching messages, especially when spreading news through Social Media, has been highlighted in several research works [32–34]. In the latter approach, different techniques were proposed for generating titles in French, and then, they were manually evaluated taking into account to what extent they were relevant, but also catchy. Focusing on the fact that companies can better exploit Twitter for distributing information in an effective manner with the help of automatic tools, such as natural language tweet generation, the research work presented in [10] analyses the way a tweet is created in B2B and B2C marketers, showing that there are differences between them that could influence in a company's presence and reputation. This analysis focus on a specific scenario, where features, such as brand names, product names, emotional language, or use of hashtags and links are studied.

The idea of our research work could be related to this one, but with the difference that we are more oriented to capture how users perceive tweets as far as the information contained and the interest they produce, in order to analyse the linguistic features that make good tweets distinguishable from bad ones. The findings of our research could benefit companies in providing tools to help generate informative and interesting information in an (semi-)automatic manner to catch customers' attention, and therefore, increase the popularity and reputation of the brand and the products associated.

3. Methodology and research questions

The methodology proposed in this research is based on descriptive statistical analysis. On the one hand, our purpose is to understand the general and most relevant properties of the dataset, and study the users' perception towards the informativeness and interestingness of a tweet (either manual or automatically generated). On the other hand, a deep linguistic analysis involving lexical, syntactic and semantic features is also conducted using different subsets of tweets, that will allow us to extract and identify the best linguistic features. The interesting aspect of this analysis is that we may be able to distinguish between good and not so good tweets, and thus, this information can be used for improving the generation of tweets.

In the next sections, the datasets, resources and tools employed are first explained (Section 3.1), and then, the research questions that are studied within this research were outlined (Section 3.2).

[5] http://www-nlpir.nist.gov/projects/duc/.

3.1. Datasets, resources and tools

The scenario and domain chosen for this research is newswire, and consequently, the data used are tweets generated from single news. The reason for choosing this domain is because at this stage of the research we are more interested in being able to characterise potential interesting informative and well-written tweets, and not taking into account highly informal tweets. As it was said in Section 2, the informality of the Web 2.0 may pose a problem to the existing NLP tools, decreasing their performance, and therefore, the quality of the results.

Specifically, as dataset we took advantage of the generated tweet collection described in [7]. This collection of tweets was generated using seven text summarisation approaches,[6] capable of producing tweets in English and Spanish from a random sample of 201 single-document news (100 news for Spanish and 101 news for English). Additionally, the original tweets associated with each news were also included in the dataset, since they were manually generated. In total, our dataset contained 1407[7] and 201[8] automatic and manually generated tweets, respectively. For both languages, the source news from which the tweets were generated were randomly chosen among the most popular news for a 10-day period in different newswire sites, such as *BBC*, or *The Guardian* for English, and *El País* or *El Mundo* for Spanish.

Regarding the other necessary resources and tools, Freeling (version 3.0) linguistic analyser [35] and IBM SPSS Statistics software (version 20) were used. The former was employed for carrying out the multilingual linguistic analysis, which comprised the identification and extraction of the lexical, syntactic and semantic linguistic features contained in the tweet collection, whereas the latter offers great capabilities to analyse the data from a statistical perspective, and therefore it was used for performing all the statistical analysis conducted in our research.

3.2. Research questions

As it was mentioned in Section 1, to conduct this research study, we analyse the dataset of tweets according to these questions:

1 ***Do users equally perceive manually and automatically generated tweets?*** Taking as a starting point the dataset of 1608 tweets, a user evaluation with respect to the aspects of informativeness and interestingness to analyse the preferences of the users, is conducted.
2 ***What linguistic features should or should not have a tweet in order to be informative, as well as interesting for the user?*** For answering this question we will take as a basis the findings obtained from the previous question. This may be the most relevant and novel contribution of this research, since it will provide us with an idea of the specific linguistic characteristics that good and bad tweets have, thus differentiating ones from the others, and also allowing us to analyse this issue from a multilingual perspective (English and Spanish).

The experiments, analysis and evaluation conducted for each of the questions is explained in the next sections in the same order they were formulated.

[6] For more detail on the text summarisation approaches employed, see [7].
[7] 100 generated tweets for Spanish × 7 summarisers + 101 generated tweets for English × 7 summarisers.
[8] 100 manual generated tweets for Spanish + 101 manual generated tweets for English.

4. Users' perception on manual and automatic generated tweets

In order to assess user's perception of the natural language generated tweets, a user study was conducted. The objective of this study was to experiment with real data and users who could receive the information through Twitter. In the evaluation, they had access to the full source news, from which the tweet was produced. The evaluation was carried out by 16 Spanish native users (6 women and 9 men between 25 and 35 years old) who were also fluent in English (having at least a B2 level according to the Common European Framework of Reference for Languages[9]).

In particular, each tweet was manually evaluated by two users according to a 3-Level Likert scale (*1 = strongly disagree; 2 = neither agree nor disagree (neutral); 3 = strongly agree*), without knowing how the tweet was generated. The use of this type of Likert scale was appropriate for our experiments [36], having been already employed for manually evaluating the output of natural language generation approaches [37,38].

The key aspects evaluated were: *informativeness* and *interestingness*. Informativeness aims to determine whether the tweet by itself provided a clear idea of the topic of the source document from which it was generated (i.e., the amount of useful information a tweet may contain). As reported in [39], there are no special studies regarding human judgement on text informativeness; however, it is a common evaluation criterion in the *INEX Tweet Contextualization* task at CLEF [40,12,13].

However, although tweets may contain valuable information, many may be not interesting to users, and finding and recommending tweets that are of potential interest to users from a large volume of tweets is a crucial but challenging task [14], even some attempts have been done in order to detect this criterion automatically [11]. In our experiment, the assessment of the interestingness aimed to capture to what extent the user's attention was drawn by the way the tweet was generated, and whether they would be curious or not in knowing more about the information provided in the tweet (e.g., by reading the whole source document or looking for more information). Specifically, two questions to rate each of these aspects were defined in our evaluation framework:

- *Informativeness*: When reading the tweet, does it provide enough information to know what the tweet is about? That is, after reading the tweet, will you be able to identify the topic of the news from which it was generated in a clear and easy way?
- *Interestingness*: Is the tweet interesting enough to catch your attention? That is, after reading it, are you curious and would you like to know and read more about the topic mentioned in it?

The reason for deciding on these two variables was due to the fact that in our long-term goal of automatically generating tweets, instead of focusing only on relevance, we want to seek for interestingness, as well. This manner an added-value to the information shown will be provided.

Since a manual evaluation is conducted, the background knowledge and interests of the users may influence on the assessment of the tweets, being reflected in the results. Despite the inherent subjectivity of the process, we believe that the positive issue is to work with real data and users in a real context, and carry out a pilot testing, so we can evaluate and analyse if the automatic generation of tweets from a general perspective and, more specifically, their quality could be potentially useful and feasible for the society.

[9] http://www.coe.int/t/dg4/linguistic/cadre1_en.asp.

Table 1
Descriptive statistics (mean and mode) for the generated tweets according to the 3-Level Likert scale (*1 = strongly disagree; 2 = neither agree nor disagree (neutral); 3 = strongly agree*).

Criteria	English		Spanish	
	Manual	Automatic	Manual	Automatic
Informativeness (mean)	2.35	2.12	2.41	1.99
Interestingness (mean)	1.89	2.04	1.87	1.76
Informativeness (mode)	3	3	3	2
Interestingness (mode)	2	3	2	1

Once our user study was conducted, we first extracted some descriptive statistics using IBM SPSS Statistics software package that are shown in Table 1.

The results obtained showed that for English, manual generated tweets obtained on average 2.35 and 1.89 for the informativeness and interestingness, respectively, whereas the same average values for automatic tweets were 2.12 and 2.04. Analysing the average values comparing manual and automatic tweets, it is interesting to note that as fas as the informativeness is concerned, although manual tweets score higher, the automatic summaries perform above 2, thus indicating that there are several tweets that are individually scored with a 3 (indeed, the mode for informativeness is 3). This finding shows the appropriateness of automatic text summarisation techniques for generating informative tweets. It also seems that users found it more interesting the automatic tweets rather than manual ones (2.04 vs. 1.89).

In the case of Spanish, the differences between manual and automatic tweets are greater both with respect to informativeness and interestingness. The average values obtained for manual and automatic tweets were: 2.41 vs. 1.87 (informativeness), and 1.99 vs. 1.76 (interestingness).

Although the average scores may seem low, having a look at the mode of each type of tweets, we obtained that for the informativeness criteria, the score most frequently assigned was the highest value in the Likert scale (i.e., 3 = strongly agree) for English tweets (manual and automatic) and Spanish manual tweets. In contrast, the value for the mode as far as the interestingness criterion is concerned differed across languages and types of tweets. In this respect, we would like to highlight that 3 (i.e., strongly agree) was the most frequent score assigned to English automatic tweets, and 1 (i.e., strongly disagree) was the most frequent score for Spanish automatic tweets. The differences between English and Spanish tweets may be due to the fact that, even though multilingual summarisers were used, the state of the art of NLP tools is more advanced in English, so tools in other languages may not perform as good, thus influencing negatively on the quality of the automatically generated tweets. Another possible reason could be the way in which tweets were generated from a linguistic point of view. We will analyse this issue in more detail in Section 5.

Table 2 shows the results for agreement between assessors computed using the Cohen's Kappa [41]. Regarding the results obtained, in general the agreement is poor. More specifically, and with respect to the interpretation of the scores [42], we got a slight agreement for the interestingness criterion for all the tweets, except for the automatic tweets in English; we obtained a fair

Table 2
Kappa scores for informativeness and interestingness.

Criteria	English		Spanish	
	Manual	Automatic	Manual	Automatic
Informativeness	22%	26%	40%	28%
Interestingness	11%	24%	16%	15%

Table 3
Percentage of tweets in which the assessors agreed. (*Informativeness/Interestingness-OK = tweets rated with the value 3 in the Likert scale; Informativeness/Interestingness-NO_OK = tweets rated with the value 1 in the Likert scale.*)

Criteria	English		Spanish	
	Manual	Automatic	Manual	**Automatic**
Informativeness-OK	29.7%	26.4%	50.0%	27.4%
Informativeness-NO_OK	5.9%	15.3%	14.0%	21.6%
Interestingness-OK	16.8%	21.5%	13.0%	10.4%
Interestingness-NO_OK	21.8%	17.8%	15.0%	15.0%

agreement for the remaining types of tweets. The highest inter-rater agreement was obtained for the Spanish manually generated tweets (40%).

This poor agreement was expected, since the task of manually evaluate different types of tweets regarding the proposed criteria (informativeness and interestingness) involves a high degree of subjectivity, and therefore, it is very difficult that two users have the same opinion for the same tweet. This is also confirmed when evaluating manual tweets for the informativeness criteria, for which there is not a substantial agreement. For this criterion, although we expected a higher agreement, it may have occurred that the original news could talk about different subtopics, and depending on which of them were considered most relevant by the users, the assessment might have been also influenced. For the interestingness criterion, the interests of users will influence their evaluation. For instance, if the tweet is about sports, and the user is not keen on that, the tweet will probably get a lower score.

Despite the subjectivity of the evaluation process, we would like to note that the Kappa scores could have been also affected by rare observations (e.g., ratings that may not be as frequent as others, even though they have been rated by the two assessors), being known as the Kappa paradox [43]. If we compared the simple inter-rater agreement with respect to the Kappa score, we obtain that for automatic English tweets there was around a 50% of agreement for both evaluated criteria, whereas Kappa values are around 25%.

Since our hypothesis is that there may be some linguistic features that can differentiate good and not so good tweets in terms of informativeness and interestingness, and given the fact that the evaluation process was very subjective, as it was shown by the Kappa inter-rater agreement, we further inspected the evaluation results, in order to analyse for how many tweets the users agreed in the fact that either they were very good (i.e., rated with 3) or very bad (i.e., rated with 1). In-between ratings (i.e., Likert scale value of 2) was discarded from this analysis, since these tweets were indifferent for the users, and therefore, their evaluation did not provide useful information. Table 3 shows the percentage of tweets falling under these categories (OK/NO_OK) for the evaluated criteria (informativeness and interestingness).

Concerning the informativeness, one can deduce that generally speaking, the tweets may help to provide an idea of what topic they are talking about, if we compared them with those ones that have been rated with the lowest value (Informativeness-OK vs. Informativeness-NO_OK). It is worth stressing the fact that for manual tweets the percentages are better in both languages, meaning that it may be easier to identify the topic in this type of tweets compared to the automatic ones, although for some cases (e.g., for English) the percentage is still low.

Regarding the interestingness, we observed a reversed trend, except for the automatic tweets generated in English. In this case, the cases in which users mostly agreed were the ones they thought that the tweets were not interesting at all. This could occur due to two issues: (i) the generated tweet is not interesting, or (ii) the user who evaluated the tweet is not keen on the topic the tweet addresses. Again, the subjectivity of the evaluation may affect the results obtained; however, since our purpose is to conduct a user study and analyse how tweets are really perceived by users, we have to assume the subjectivity involved in the process.

Given that an informative tweet may not be interesting to a user [14], we also wanted to determine the set of tweets that met both criteria at the same time, and not only one of them independently. This manner, we could analyse possible linguistic traits or features that may characterise these tweets. Therefore, we narrowed our analysis and different subsets based on the user evaluation and agreement were produced. Specifically, two subsets were obtained based on the given scores (best and worst) with two degrees of flexibility (restrictive and non-restrictive) each one. For building them, the following rules were applied:

- Subset *Best-Tweets-Restrictive*: we ensure that the two users evaluating a specific type of tweets agreed on the score. Both users assigned a tweet the highest value in the Likert scale (i.e., 3) for informativeness as well as for interestingness.
- Subset *Worst-Tweets-Restrictive*: we ensure that the two users evaluating a tweet agreed on the score, assigning them the lowest value in the Likert scale (i.e., 1) for informativeness as well as for interestingness.
- Subset *Best-Tweets-Non-Restrictive*: in this subset the agreement was slightly relaxed, and in this case, we only required that at least one of the two users scored the tweet with the highest value in the Likert scale (i.e., 3) for informativeness as well as for interestingness.
- Subset *Worst-Tweets-Non-Restrictive*: it is the same case as the previous one, but selecting those tweets that were scored the lowest in the Likert scale (i.e., 1) for informativeness as well interestingness by at least one of the two users.

The percentage of resulting tweets in each subset can be seen is Table 4. This table also shows (in brackets) the number of tweets included for each percentage out of the total tweets for each subset (707 and 700 automatic tweets for English and Spanish, respectively; and 101 and 100 for manual tweets in English and Spanish, respectively).

The results obtained show clear differences in the percentage of tweets that are selected for each language. Whereas in English, the percentage of best tweets in the restrictive and non-restrictive

Table 4
Percentage and number of tweets out of the total number of manual and automatic tweets for each language, respectively, that have been included in each subset (*Best/Worst-Tweets-Restrictive/Non-Restrictive = tweets rated with 3 or 1 (for Best and Worst, respectively) according to the Likert scale, and taking into consideration when the two assessors agreed in the rating (Restrictive) or at least one (Non-Restrictive)*).

Criteria	English		Spanish	
	Manual	Automatic	Manual	Automatic
Best-Tweets-Restrictive	7.92%(8)	11.74%(83)	6.0%(6)	5.14%(36)
Worst-Tweets-Restrictive	2.97%(3)	8.63%(61)	13.0%(13)	6.86%(48)
Best-Tweets-Non-Restrictive	35.64%(36)	40.31%(285)	41.0%(41)	26.71%(187)
Worst-Tweets-Non-Restrictive	17.82%(18)	32.39%(229)	26.0%(26)	27.86%(195)

Table 5
Examples of tweet content after users' evaluation (*EN = English; ES = Spanish*). Translations for the Spanish tweets are provided in brackets.

Good manual tweet (EN)	Gorilla genome analysis reveals new human links
Good automatic tweet (EN)	First full sequence of gorilla genome shows 96 share with humans, with close parallels in sensory perception and hearing
Bad manual tweet (EN)	Ali Dizaei: The 'copper' who refuses to go quietly – Profiles – People – The Independent
Bad automatic tweet (EN)	This trial suggests the default position should be the other way round, because most people are benefiting
Good manual tweet (ES)	El nuevo iPad, más barato que una acción de Apple (*The new iPad, cheaper than an Apple stock*)
Good automatic tweet (ES)	La consultora Gartner resalta la capacidad de Apple para darle al consumidor lo que necesita. Destaca la facilidad de uso. (*Gartner company highlights the ability of Apple to give consumers what they need. It emphasizes ease of use*)
Bad manual tweet (ES)	¡Háztelo tú mismo! (*Do it yourself!*)
Bad automatic tweet (ES)	Paso a paso. A largo plazo (*Step by step. In the long-term*)

subsets is always higher than the percentage of worst tweets for manual and automatic generated tweets, we did not obtain the same findings for the Spanish tweets. In this case, the percentage of worst tweets (manual and automatic) is higher. This only occurs when the degree of flexibility is stricter, requiring the same scoring for the assessors. In the non-restrictive subset, the percentage of best tweets for Spanish almost doubled the percentage of worst ones for the manual tweets; however, the figures for the automatic ones are very similar. Despite that this issues has to be further analysed, the difference in language may indicate that in general users find Spanish tweets worse than English ones, as it was previously stated.

Regarding the comparison between manual and automatic tweets, it is worth highlighting that for English, the percentage of automatically generated tweets that have been best scored is slightly higher than the manual tweets. As it was previously stated, despite the subjectivity that may be involved in the process, this is a positive finding, since it means that state-of-the-art summarisers systems are useful for determining relevant information, thus extracting a sentence that helps users to know what the tweet is about, and being it also interesting from a user's perspective. For Spanish though, it happens something unexpected: the percentage of worst manual tweets is equal or higher than the percentage for automatic tweets. This is interesting, since it means that the tweets generated by humans with the purpose of providing a headline of a news may not be appealing for users, and therefore other ways of generating such headlines are needed.

Examples of English and Spanish good and bad manual and automatic generated tweets extracted from the restrictive subsets are illustrated in Table 5.

Having analysed these subsets of tweets, they will be further taken into consideration for analysing the linguistic features contained in order to be capable to come up with some differences from a linguistic perspective.

5. Exhaustive linguistic analysis on manual and automatic generated tweets

In this section, given the different subsets of tweets previously categorised (Section 4), our aim is to determine whether there are any specific linguistic features that allow us to distinguish between them. Therefore, for achieving our goal, a three-step process was defined. First, the set of lexical, syntactic, and semantic features was determined; second, these features were computed over the tweets in each group using Freeling linguistic analyser; and finally, the different groups were compared. Next, these three steps are explained in detail:

- *Determining the set of features*

  Since we want to carry out an analysis as complete as possible, we define a set of linguistic features from three perspectives: lexical, syntactic, and semantic. In short, we want to know what linguistic elements could make a tweet more or less appropriate, if there were any.

  Regarding the lexical features we defined, we relied on the output of the pos-tagger and we include in this type all the information that could be extracted from it (mainly different kinds of words), as well as the number of words of tweet, and number of URLs. As far as the syntactic features is concerned, we only considered under this category features related to noun-, verb- or prepositional-phrases, extracted after performing a syntactic parsing for the tweet. Finally, semantic features were extracted after analysing the tweets using a named entity recogniser and determining the degree of polysemy of the words included, or to what extent tweets contain words with or without semantic charge (i.e., stopwords). In this case, we wanted to gather a set of features that could be representative enough to represent some semantic linguistic elements. It is worth mentioning that readability features were not taken into account, since the analysis of the difficulty of a text was out of the scope in our study. This features would have been more appropriate when one wants to generate simpler or easier tweets. It could have been also interesting to use this type of features if the tweets had grammatical errors, but this did not happen in our dataset. We rely on formal tweets coming from newswire documents (most of them headlines) or extractive sentences from the documents themselves when automatic summarisation systems were employed, and none of them contained truncated sentences.

  Table 6 provides detailed information about the features. We finally obtained a total of 56 linguistic features, differentiating between 44 lexical, 3 syntactic, and 9 semantic.

  Among all the proposed features, it is worth justifying the rationale behind features F49–F50–F51 and F54. On the one hand, concerning the types of named entities (F49, F50, F51), the types of PERSON, LOCATION and ORGANISATION were only taken into account, since they are standard entities recognised by the great majority of named entity recognisers systems [44,45]. Most systems often detect the type MISC as well, but this type was discarded in our research, because it was rather generic, thus being not useful for analysing in detail and obtaining knowledge from our data. On the other hand, the reason for determining the number of polysemic words with more than three senses (F54) was not randomly proposed. We specifically checked and computed the average number of senses in the words contained in WordNet[10] for English, and Multi-WordNet for Spanish, through the Multilingual Central Repository.[11] The average number of senses for words in English was 2.89, whereas for Spanish was 2.04. In this manner, words with more than three senses were considered, in order to determine the number of words that were above the mean according to the number of senses.

- *Computing the features*

  The previous set of features was computed for each group of tweets defined in Table 4 using Freeling linguistic analyser. We

[10] http://wordnet.princeton.edu/wordnet/man/wnstats.7WN.html#toc3.
[11] http://adimen.si.ehu.es/cgi-bin/wei/public/wei.consult.perl.

Table 6
Set of linguistic features (*Level: Lex = lexical, Syn = syntactic, and Sem = semantic*).

Level	Id	Feature	Description
Lex	F1	NumbURLS	Number of URLs/links
Lex	F2	NumbWords	Number of words
Lex	F3	AvgCharInWords	Average number of characters per word
Lex	F4	NumbSingular	Number of singular words
Lex	F5	NumbPlural	Number of plural words
Lex	F6	NumbNouns	Number of nouns
Lex	F7	NumbComNouns	Number of common nouns
Lex	F8	NumbPropNouns	Number of proper nouns
Lex	F9	NumbVerbs	Number of verbs
Lex	F10	NumbMainVerbs	Number of main verbs
Lex	F11	NumbAuxVerbs	Number of auxiliary verbs
Lex	F12	NumbVerbsP	Number of verbs in present tense
Lex	F13	NumbVerbsPP	Number of verbs in past tense
Lex	F14	NumbVerbsF	Number of verbs in future tense
Lex	F15	NumbVerbsInf	Number of verbs in infinitive form
Lex	F16	NumbVerbsPart	Number of verbs in participle form
Lex	F17	NumbVerbsGer	Number of verbs in gerund form
Lex	F18	NumbVerbsCond	Number of verbs in conditional form
Lex	F19	NumbVerbs1per	Number of verbs in first person form
Lex	F20	NumbVerbs2per	Number of verbs in second person form
Lex	F21	NumbVerbs3per	Number of verbs in third person form
Lex	F22	NumbAdj	Number of adjectives
Lex	F23	NumbQualAdj	Number of qualifying adjectives
Lex	F24	NumbOrdAdj	Number of ordinal adjectives
Lex	F25	NumbCompAdj	Number of comparative adjectives
Lex	F26	NumbSuperAdj	Number of superlative adjectives
Lex	F27	NumbAdv	Number of adverbs
Lex	F28	NumbPron	Number of pronouns
Lex	F29	NumbPersPron	Number of personal pronouns
Lex	F30	NumbDemPron	Number of demonstrative pronouns
Lex	F31	NumbPosPron	Number of possessive pronouns
Lex	F32	NumbIndefPron	Number of indefinite pronouns
Lex	F33	NumbInterPron	Number of interrogative pronouns
Lex	F34	NumbRelPron	Number of relative pronouns
Lex	F35	NumbExclPron	Number of exclamative pronouns
Lex	F36	NumbDeterm	Number of determiners
Lex	F37	NumbConj	Number of conjunctions
Lex	F38	NumbCoordConj	Number of coordinated conjunctions
Lex	F39	NumbSuborConj	Number of subordinated conjunctions
Lex	F40	NumbInterj	Number of interjections
Lex	F41	NumbPrep	Number of prepositions
Lex	F42	NumbPunct	Number of punctuation marks
Lex	F43	NumbNumerals	Number of numerals
Lex	F44	NumbDate	Number of date and time expressions
Syn	F45	NumbNP	Number of noun phrases
Syn	F46	NumbVP	Number of verb phrases
Syn	F47	NumbPP	Number of prepositional phrases
Sem	F48	NumbNER	Number of named entities (NER)
Sem	F49	NumbNER-PER	Number of PERSON named entities
Sem	F50	NumbNER-LOC	Number of LOCATION named entities
Sem	F51	NumbNER-ORG	Number of ORGANISATION named entities
Sem	F52	NumbMultiWords	Number of multiwords
Sem	F53	NumbPolysemic	Number of polysemic words
Sem	F54	NumbPolysemic-3	Number of polysemic words with >3 senses
Sem	F55	NumbMonosemyc	Number of monosemyc words
Sem	F56	NumbSemWords	Number of words with semantic charge

decided to use Freeling, since it provides linguistic analysis at a lexical, syntactic and semantic level for both English and Spanish languages.

Due to the inherent nature of tweets, NLP processes may have problems in dealing with Twitter texts, especially because: (i) they are too short, and therefore, there is not context associated; and, (ii) they are informal, thus being generally ill-formed texts (e.g "*I'm liking this It's on Us thing from Lloyds Bank:) #hellofreemoney*"). However, it is worth mentioning that although it is possible to think that Freeling may have problems when processing tweets, in our case, the tweets we deal with do not contain any informal language, because they are derived from newswire and therefore, Freeling is not affected by this phenomenon.

Once the different features were computed, each tweet was considered as a vector of linguistic features, allowing the analysis and comparison of the different groups, which is next explained.

- *Comparing features*

For analysing the tweets from a linguistic point of view, the IBM SPSS Statistics software package was employed. In particular, we used the Mann–Whitney *U* non-parametric test. We opted for this type of test, since our sample was not big enough (some sample sizes had less than 30 elements) to assume and ensure normality [46,47], and therefore, we could not apply commonly used parametric tests, such as *T*-test or Anova.

Taking into account the previously mentioned subset of tweets (defined at the end of Section 4, a multi-dimensional

Table 7
Tests and comparisons performed (*EN = English; ES = Spanish; R = restrictive; NR = non-restrictive*).

(Test number) Comparison groups
(1) Best EN Manual NR vs. Best EN Automatic NR; (2) Best EN Manual R vs. Best EN Automatic R; (3) Worst EN Model NR vs. Worst EN Automatic NR; (4) Worst EN Model R vs. Worst EN Automatic R; (5) Best EN Manual NR vs. Worst EN Manual NR; (6) Best EN Manual R vs. Worst EN Manual R; (7) Best EN Automatic NR vs. Worst EN Automatic NR; (8) Best EN Automatic R vs. Worst EN Automatic R; (9) Best ES Manual NR vs. Best EN Automatic NR; (10) Best ES Manual R vs. Best EN Automatic R; (11) Worst EN Manual NR vs. Worst ES Automatic NR; (12) Worst EN Manual R vs. Worst ES Automatic R; (13) Best ES Manual NR vs. Worst ES Manual NR; (14) Best ES Manual R vs. Worst ES Manual R; (15) Best ES Automatic NR vs. Worst ES Automatic NR; (16) Best ES Automatic R vs. Worst ES Automatic R; (17) Best EN Manual NR vs. Best EN Manual R; (18) Best EN Automatic NR vs. Best EN Automatic R; (19) Worst EN Manual NR vs. Worst EN Manual R; (20) Worst EN Automatic NR vs. Worst EN Automatic R; (21) Best ES Manual NR vs. Best ES Manual R; (22) Best ES Automatic NR vs. Best ES Automatic R; (23) Worst ES Manual NR vs. Worst ES Manual R; (24) Worst ES Automatic NR vs. Worst ES Automatic R; (25) Best EN Manual NR vs. Best ES Manual NR; (26) Best EN Manual R vs. Best ES Manual R; (27) Best EN Automatic NR vs. Best ES Automatic NR; (28) Best EN Automatic R vs. Best ES Automatic R; (29) Worst EN Manual NR vs. Worst ES Manual NR; (30) Worst EN Manual R vs. Worst ES Manual R; (31) Worst EN Automatic NR vs. Worst ES Automatic NR; (32) Worst EN Automatic R vs. Worst ES Automatic R

analysis was established with these four criteria: (i) language (English vs. Spanish); (ii) degree of flexibility[12] (restrictive vs. non-restrictive); (iii) method for generating the tweet (manual vs. automatic); and (iv) scoring (best vs. worst). We also include the comparisons between restrictive and non-restrictive sets, and English and Spanish tweets for several reasons. On the one hand, although the set of tweets in the restrictive set is already included in the non-restrictive, the comparison of these tweets is also interesting to know whether there are significant differences between them. If so, this will mean that the part of tweets that are not identical contains some characteristics that would be worth further analyse. On the other hand, concerning the languages, the direct comparison between English and Spanish is also interesting to confirm if there are any differences in the way tweets are written with respect to their lexical, syntactic or semantic features.

Based on the aforementioned dimensions, we compared the tweets as linguistic vectors in the different groups taking two at a time, since a priori we guess that there will be differences between them. In total, we run 32 Mann–Whitney *U* tests, setting one of the criteria each time (e.g., the language) and varying the remaining ones (e.g., degree of flexibility, method, and scoring). Table 7 shows the tests performed. This table will be useful to know which groups exhibits statistical significant differences for which features and type of features. Also, based on the results, they will be used to decide which tests are the most relevant ones to be further analysed in more detail.

Only the combinations that were most interesting for our research were compared and analysed, and in this section we only show the most relevant findings and results.[13] For accounting statistically significant differences, we consider a 95% confidence interval, so when the significance value is equal or lower than 0.05, this means that there are significant differences between the groups of tweets compared.

Three types of analysis were conducted for the results obtained: (i) a general analysis was carried out over all the tests and features; (ii) a feature-based analysis, accounting for the features involved in each test; and (iii) a specific analysis where the behaviour of some features for some particular tests are compared and discussed in order to determine which could be the most representative features for each type of tweets. Next, we provide more detail about the analysis performed and the most relevant findings obtained.

– *General analysis*. We first analysed the average number of tests in which the different types of features (lexical, syntactic, and semantic) showed to be statistically different. In this manner, we can obtain a general idea of the type of features that could contribute to make a tweet different. Table 8 shows the results. As it can be seen, semantic and syntactic features prevail over the lexical ones, although a higher number of lexical features was defined. In particular, comparing the number of features for each linguistic level to the number of tests in which that features were significantly different, we observe that syntactic features, followed by semantic ones seem to have more power to distinguish between tweets. Although 3 features were only identified for this type, they show significant differences in more than half of the tests performed. Semantic features are the type of features contributing the second most to the differences between tweets and finally, lexical features seem to show less differences, despite the number of them is larger than the other type of features.

Table 8
Linguistic influence of the features over the tweets (percentage of linguistic features that are significant in the tests performed).

Linguistic feature	Influence
Lexical	20.69%
Syntactic	43.12%
Semantic	36.20%

None of the tests performed (i.e., any comparison between two groups of tweets) showed significant differences in all the features investigated. As far as the lexical features is concerned, only test 27 (i.e., comparison between best tweets automatically generated in English and Spanish) exhibit differences in the 72% of the lexical features. This is expected since, as we previously stated, English and Spanish language have different origins, and it is logical that tweets differ in most of the lexical features at least. Regarding syntactic features, we found significant differences for all of them in several tests. Again, it occurs that most of these tests show that we do find syntactic differences when comparing tweets in different languages, which is logical (test 12, 27, 28, and 31), but also when comparing best and worst tweets regardless the way they were generated and the language (test 8, 13, 14, 15, 16). Finally, concerning semantic features, only test 7 shows differences in all the semantic features. This is a very interesting finding, since this test compares good and bad automatically generated tweets in English, and it means that the language employed is different from a semantic perspective. Further on, we will provide a more detailed analysis of the differences.

On the contrary, we also analysed the groups of tweets that did not show any differences at all (tests 17 and 21), or marginal ones (i.e., having at least 95% of the features without statistical differences) found in tests 6, 10, 19, 22. Most of these tests concern the comparison between restrictive and non-restrictive datasets, so in these cases, part of the tweets

[12] This takes into account whether we require that the tweet had a complete agreement in its indicativeness and interestingness by the two uses who evaluated it or not, explained in Section 4.

[13] Due to the big dimensions of the table, the whole results (significance value) of the Mann–Whitney *U* for all tests and features can be accessed at: http://goo.gl/mKSe88http://goo.gl/mKSe88http://goo.gl/mKSe88, and it is also provided as supplementary data.

Table 9
Linguistic unique features for distinguishing good and bad manual and automatic tweets in the restrictive tweets datasets (*EN = English; ES = Spanish*).

	Best manual vs. automatic	Worst manual vs. automatic	Best vs. worst manual	Best vs. worst automatic
EN	F2, F10, F37, F45, F47, F56 (test 2)	F4, F6, F7, F33, F42, F48, F49, F53, F56 (test 4)	F42, F49 (test 6)	F2, F3, F4, F5, F6, F7, F9, F10, F12, F13, F15, F17, F21, F22, F26, F31, F37, F42, F45, F46, F47, F48, F50, F51, F53, F54, F55, F56 (test 8)
ES	F33 (test 10)	F1, F2, F3, F4, F5, F6, F7, F8, F9, F10, F12, F21, F22, F23, F36, F41, F42, F43, F45, F46, F47, F48, F49, F53, F54, F55, F56 (test 12)	F1, F2, F3, F4, F5, F6, F7, F8, F9, F10, F12, F13, F15, F21, F22, F23, F36, F41, F45, F46, F47, F48, F50, F51, F53, F54, F55, F56 (test 14)	F2, F4, F6, F7, F8, F9, F10, F12, F15, F19, F21, F22, F24, F28, F29, F36, F41, F43, F45 (test 16)

included in the restrictive subset could be also take part in the non-restrictive ones. However, special attention is worth paying to tests 6, and 10. On the one hand, test 10 indicates that there is no difference between the best manual and automatic tweets for English, except for F33 (number of interrogative pronouns). In this manner, we can deduce that to some extent automatic summarisers can also generate very good tweets. On the other hand, test 6 indicates that there are only linguistic differences between the best and worst manually generated tweets for F42 (punctuation marks) and F49 (person named entities). Manually generated tweets either good or bad may be correct from a linguistic point of view, and in this case, the subjectivity involved in the evaluation process could have influenced on the results obtained.

– *Feature-based analysis*. Focusing on the features, the first issue that drew our attention was that five of them did not exhibit any significant differences for any test. This occurred for the lexical features: number of verbs in second person form (F20); number of demonstrative pronouns (F30); number of exclamative pronouns (F35); number of coordinated conjunctions (F38); number of subordinated conjunctions (F39); and number of interjections (F40). Moreover, we checked the individual values obtained for each tweet, and none of the tweets in our dataset contained three of them (F35, F38, and F39). This may be explained by the fact that they the presence of exclamative pronouns may be rare in formal tweets, as well as the presence of coordinated and subordinated conjunctions, due to the length constraints. Tweets are normally very short (only 140 characters), and thus, they normally contain a single-sentence, so these linguistic elements may not normally appear. However, analysing the number of conjunctions from a general perspective (F37), we observe that this feature is indeed present in some of the tweets, and some groups differ significantly in the number of conjunctions (e.g., groups in test 1). This may be due to the fact that: (i) conjunctions appears for joining other types of linguistic elements which are not clauses; or (ii) some tagging errors may be propagated by the linguistic analysis performed with Freeling.

Continuing with our analysis, we found other features that are only significant in three tests at most. Specifically, the number of conditional verbs (F18) and comparative adjectives (F25) are only significant in test 27; indefinite pronouns (F32) is significant in tests 27 and 31; and the feature corresponding to the number of verbs in future tense (F14) and verbs in first person (F19) are significant in tests 15, 16 and 27; and 7, 16, and 27, respectively. Analysing these features in detail, these are again lexical features, and according to the types of test, we observe that these differences appear when we mostly compare tweets across English and Spanish (e.g., test 27 and 31), and not in tests within the same language. Since we compare languages of different nature (English vs. Spanish), our intuition behind this is related the way a language is used. Both languages differ in lexical, syntactic and semantic aspects, so this may determine how frequent or event the presence of some of the features [48–50].

On the contrary, we have identified a set of features that showed to be significantly different for a high number of tests. In this case, the most frequent feature is the number of words with semantic charge (F56) that it is significant in 75% of the tests performed (24 tests out of 32). The remaining 5 most frequent features shown in descending order by the number of tests in which this feature is significant are: the number of polysemic words (F53) in 22 tests; the number of words in the tweet (F2) in 21 tests; the number of words in singular (F4) in 21 tests; the number of common nouns (F7) in 21 tests; and the number of noun phrases (F45) in 20 tests. In this respect we observe that general differences between tweets, regardless the language and how they were generated are: the number of words (F2), common nouns (F7) and words in singular form (F4) they contained, as far as the lexical features is concerned; the number of noun phrases (F45) from a syntactic perspective (this may be related to the finding for the lexical feature F7); and finally, the number of words with semantic charge (F56), and the number of polysemic words (F53), with respect to the semantic level.

– *Determining the most representative features*. Now, we would like to provide a more detailed analysis, focusing on the features that may be unique and decisive for contributing to such differences. Here, we limit our analysis to the restrictive datasets only, since it contains the set of tweets in which there was a fully agreement between users with respect to the informativeness and interestingness of a tweet. Taking into account the tests performed for these sets, we studied the features characterising the manual and automatic tweets with respect to the best and worst sets. Table 9 shows the results obtained.

From the previous table, we can see some common features among the groups analysed, as well as some features that are unique for each group. Focusing on English tweets, we observe that all the features that differentiate manual vs. automatic best tweets (test 2) are also present in the comparison between best vs. worst automatic tweets (test 8). This may indicate that such features are the ones that characterise automatic best tweets. Considering these sets we do not take into account the number of words with semantic charge (F56), since it appears in test 2 and test 4, so it does not contribute to differentiate best and worst tweets. Analogously, most of the features contained in test 4, except the number of interrogative pronouns (F33) and the PERSON type named entities (F49) are also present in test 8, indicating that these features characterise worst automatic tweets. Punctuation marks (F42) would be the feature that characterise worst manual and automatic tweets, whereas the PERSON type named entities only does so for manual ones.

Table 10
Average values obtained for relevant features in the different subsets of restrictive tweets (*B = best; W = worst; Auto = automatic*).

Lang.	Feature (desc)	B-manual	B-auto	W-manual	W-auto
EN	F2(Words)	11.25	17.82	15.33	7.68
	F4 (Singular)	2.75	3.06	2.33	0.57
	F6 (Nouns)	3.75	4.31	3	0.9
	F7 (Common Nouns)	3.75	4.31	3	0.9
	F10 (MainVerbs)	0.875	1.73	1	0.78
	F37 (Conj)	0.75	2.29	0	0.8
	F42 (Punct)	1.375	1.88	4.33	1.46
	F45 (NP)	0.875	2.33	0.67	0.88
	F47 (PP)	0.875	1.94	0.33	0.6
	F48 (NER)	1.125	1.37	2.33	0.54
	F49 (NER-PER)	0.125	0.26	1.33	0.25
ES	F1 (URLs)	0	0	0.92	0
	F2 (Words)	11.17	13.84	1.85	7.87
	F3 (AvgCharInWords)	5	5.37	59.31	4.87
	F4 (Singular)	4.17	5.63	0.23	2.66
	F5 (Plural)	1.5	1.84	0	1.07
	F6 (Nouns)	3.17	3.47	0	2.35
	F7 (CommonNouns)	2	2.49	0	1.72
	F8 (ProperNouns)	1.17	0.98	0	0.63
	F10 (MainVerbs)	1.33	1.51	0.08	0.75
	F12 (VerbsPresent)	0.5	1.05	0	0.45
	F21 (Verbs3Per)	0.67	1.28	0	0.47
	F22 (Adj)	0.83	0.74	0	0.28
	F33 (InterPron)	0.17	0	0	0.01
	F41 (Prep)	1.83	2.19	0	0.9
	F45 (NP)	2.67	3.47	0	2.25

Regarding Spanish tweets, we observed that interrogative pronouns (F33) distinguishes between manual and automatic tweets as far as how good they are (test 10). However, we did not obtain any conclusive results concerning the decisive features for characterising both manual and automatic best tweets. When it comes to identifying a bad tweet, it seems that for manually generated tweets is much easier to identify a bad tweet in Spanish than in English, since we have a higher number of unique and decisive features common for worst manual and automatic tweets: number of URLs (F1); number of words (F2); average number of characters per word (F3); number of words in singular (F4); number of words in plural (F5); number of nouns (F6); number of common nouns (F7); number of proper nouns (F8); number of main verbs (F10); number of verbs in present tense (F12); number of verbs in third person (F21); number of adjectives (F22); number of prepositions (F41); and F45 (number of noun phrases). Except the latter, all these features are lexical ones.

Distinguishing between manual and automatic tweets, worst manual tweets differs from the best ones in the number of URLs they contain (F1); the number of verb phrases (F46); the number of prepositional phrase (F47); the number of named entities (F48); the number of polysemic words (F53); the number of polysemic words with more than 3 senses (F54); the number of monosemyc words (F55); and the number of words with semantic charge (F56). As it can be seen, most of the differences are exhibited for syntactic and semantic features, except the number of URLs.

Moreover, automatically generated best and worst tweets differs in the number of prepositions (F41), and the number of numerals (F43).

After the analysis reported, as a general finding it seems that the number of URLs (F1) characterise worst tweets in general, and manual tweets in particular. It seems that when a tweet contains an external link with very little additional information or without any, it is negatively considered by users, since one cannot have an idea of the content of such link until it is opened in the browser.

In order to check that our previous intuitions are correct, and whether the proposed selected features are predominant in the best or worst tweets, we analysed and compared the exact value obtained for the most relevant features found in the different types of tweets. Table 10 shows the figures obtained.

Analysing the meaning behind the selected features, and taking into account the previous general discussion, our starting point for analysing this table is that, for English, the number of words (F2); the number of main verbs (F10); the number of conjunctions (F37); the number of noun phrases; and the number of prepositional phrases (F47) may represent in a unique way good tweets, whereas number of words in singular (F4); number of nouns (F6); number of common nouns (F7); number of punctuation marks (F42); number of named entities (F48); and number of PERSON named entities (F49) are characteristic of bad tweets. First of all, the length of a tweet appears to influence on users, being longer tweets better than shorter ones. The number of main verbs as well as the number of conjunctions is higher than 1, so this means that a good tweet maybe covering more than one idea or topic. Finally, regarding syntactic linguistic aspects, we found out that the number of noun phrases and prepositional phrases is also higher in good tweets than in bad ones, so this means that the more information provided for the topic of the tweet, the better.

In the case of bad automatic tweets, the presence of key elements in tweets, such as nouns (or common nouns), words in singular, and named entities was insufficient, so this reflects the importance of considering lexical and semantic elements when generating tweets automatically. The number of named entities (F48) indicates that it would be better to focus on a

named entity (when possible) than not consider them. Similar conclusions can be drawn for nouns. The results indicate that if a tweet do not consider this type of words, it has more chances not to be informative or interesting. An interesting finding is related to the number of punctuation marks (F42), where exceeding the number of punctuation marks in a tweet may have a negative impact, as it occurs for the worst manual tweet group.

For Spanish, as we previously stated, we come up with a set of features that were able to characterise bad tweets, regardless the way they were generated (manual or automatic ones). These features were: number of URLs (F1); number of words (F2); average number of characters per word (F3); number of words in singular (F4) or in plural (F5); number of nouns (F6); number of common (F7) or proper nouns (F8); number of main verbs (F10); number of verbs in present tense (F12); number of verbs in third person (F21); number of adjectives (F22); number of prepositions (F41); and number of noun phrases (F45). Because it was clear in this language that bad tweets could be easier to identify than for English, we analysed the content of the individual tweets in this group, and we realised that for Spanish, most of the worst tweets contained only a link pointing to an external information source. This is why the number of external links a tweet had (F1) was considered one of the features representing bad tweets, and this also explains the high number of characters per word on average for worst manual tweets that it is far too large with respect to the values obtained for the remaining sets (value for F3: 59.31, because the whole URL was considered as a single word). Regarding the number of words, the results are in line with the English ones, and again we obtain that too short tweets may lack of informativeness and interestingness. Concerning the types of linguistic elements (nouns, verbs, etc.), it is important to observe that worst tweets include them to a much lesser extent than best tweets. This happens for instance, for nouns (and therefore noun phrases), main verbs, adjectives or prepositions.

Turning to characterising best tweets in Spanish, we could only identify one feature distinguishing good manual tweets. This feature represents interrogative pronouns (F33), and it may indicate that the use of this type of pronouns is useful when generating tweets from news (who, what, where, etc.), as it happens for the manual tweets analysed.

6. Conclusion and future work

In this article, an in-depth analysis concerning the users' perception for natural language generated tweets was carried out. These tweets where generated either manually or automatically, through state-of-the-art multilingual summarisers capable of producing summaries as short as tweets.

Our analysis focused on two issues that have great importance in tweets and should be taken into account in a joint manner (informativeness and interestingness), since they may influence on the impact a tweet will have on users. Moreover, our research was guided by two research questions: (1) do users equally perceive the tweets that have been manually generated in comparison to automatic generated ones?; and (2) what linguistic features should or should not a good tweet have in order to be interesting, as well as informative for the user?. For answering these questions, the proposed methodology was based on a descriptive statistical analysis, on the one hand determining to what extent the tweets in our dataset were informative and interesting, and on the other hand, conducting an additional linguistic analysis in order to extract and identify the features (lexical, syntactic and semantic) that will improve the automatic generation of tweets.

The results of this paper show that concerning the answer to the first research question, it is possible to generate good automatic natural language tweets, that are informative as well as interesting, and could be an alternative to manual generated tweets. Regarding to the second research question, it was proved that it is possible to distinguish and characterise good and bad tweets based on different linguistic features that are not present in other types of tweets. As far as the linguistic analysis is concerned, the most relevant findings are that for English, good tweets are characterised by the number of words, main verbs, conjunctions, noun-phrases and prepositional phrases, whereas bad tweets can be characterised by the number of words in singular, nouns and common nouns, punctuation marks, named entities and person named entities. In contrast, for Spanish, we found out more linguistic features for representing bad tweets, stressing the fact that when a tweet is generated only by a link, this is not a good strategy to follow, and the low presence of nouns, noun phrases, main verbs, adjectives or prepositions may have also a negative influence both in the informativeness and interestingness of the tweet.

It is worth mentioning that although our experiments were conducted in the context of Twitter, the results and findings obtained could be extrapolated and applied to the task of generating short messages, headlines, key ideas, etc., taking into account relevance but also interestingness.

The overall long-term goal of our research work, given our experience and background in text summarisation, is the automatic generation of quality and efficient natural language tweets from newswire documents using linguistic information at different levels. This task can help companies to automate marketing-related issues. For instance, by developing approaches that support companies when it comes to writing and distributing effective messages through Social Networks (e.g., informative as well as interesting tweets), or marketers that want their audiences to engage with their brand messages.

From the findings obtained in this pilot evaluation, we would like to broad the experiments using a higher number of tweets and users, employing better data collection methods, such as the ones suggested for instance in [51,52] or [53]. Since manual evaluation is a very costly and time-consuming task, we plan to precisely define an improved evaluation environment, and use crowdsourcing strategies (e.g., Crowdflower[14]) specifying the appropriate requirements to collect task-committed users, so that the experiments could be replicated in a wider context, and other variables could be analysed. Using these type of platforms, we will ensure that native users in the language of the tweet assess them (e.g., English native users will evaluate English tweets). Moreover, we would like to further investigate, develop and test an automatic model associated to the set of most relevant features discovered, in order to be able to extract relevant content and classify messages into good and bad, as well as the predictive power of each of the features. The inclusion of other type of linguistic features (e.g., discursive or concerning the readability) would be also another issue to investigate in the short term, since they could provide more deep knowledge about interesting aspects of the tweet, such as its purpose, coherence, etc., or how accessible and inclusive a tweet would be. Finally, the linguistic analysis for each evaluated criteria (informativeness vs. interestingness) from an independent manner would be useful for determining whether part of our findings are confirmed or not, and for deciding which linguistic elements may be compatible or contradictory between them.

[14] http://www.crowdflower.com/.

Acknowledgements

This research work has been partially funded by the University of Alicante, Generalitat Valenciana, Spanish Government and the European Commission through the projects, "Tratamiento inteligente de la información para la ayuda a la toma de decisiones" (GRE12-44), "Explotación y tratamiento de la información disponible en Internet para la anotación y generación de textos adaptados al usuario" (GRE13-15), DIIM2.0 (PROMETEOII/2014/001), ATTOS (TIN2012-38536-C03-03), LEGOLANG-UAGE (TIN2012-31224), and SAM (FP7-611312).

Additionally, I would like to thank you all the users who contribute to the evaluation and especially Miguel Ángel Guerrero and Raúl Bernabeu for their help and support in the statistical analysis of the article.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.compind.2015.10.010.

References

[1] A. Fensel, I. Toma, J.M. García, I. Stavrakantonakis, D. Fensel, Enabling customers engagement and collaboration for small and medium-sized enterprises in ubiquitous multi-channel ecosystems, Comput. Ind. 65 (5) (2014) 891–904, Special Issue: New trends on E-Procurement applying Semantic Technologies.
[2] B. Grabot, A. Mayere, F. Lauroua, R. Houe, ERP 2.0, what for and how? Comput. Ind. 65 (6) (2014) 976–1000.
[3] M. Banbersta, The Success Factors of the Social Network Sites Twitter, Tech. Rep., Corssmedialab, Utrecht University of Applied Sciences, 2010.
[4] B. Jansen, M. Zhang, K. Sobel, A. Chowdury, Twitter power: tweets as electronic word of mouth, J. Am. Soc. Inf. Sci. 60 (11) (2009) 2169–2188.
[5] S.E. Cho, H.W. Park, Who are dominant communicators on twitter? A study of Korean twitter users, Int. J. Contents 9 (1) (2013) 49–59.
[6] K. Ganesan, C. Zhai, E. Viegas, Micropinion generation: an unsupervised approach to generating ultra-concise summaries of opinions, in: Proceedings of the 21st International Conference on World Wide Web, ACM, New York, NY, USA, 2012, pp. 869–878.
[7] E. Lloret, M. Palomar, Towards automatic tweet generation: a comparative study from the text summarization perspective in the journalism genre, Expert Syst. Appl. 40 (16) (2013) 6624–6630.
[8] B. Dorr, D. Zajic, R. Schwartz, Hedge trimmer: a parse-and-trim approach to headline generation, in: Proceedings of the HLT-NAACL 03 on Text Summarization Workshop – Volume 5, HLT-NAACL-DUC'03, Association for Computational Linguistics, Stroudsburg, PA, USA, 2003, pp. 1–8.
[9] C. Lopez, V. Prince, M. Roche, Just title it! (by an online application), in: Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Avignon, France, 2012, pp. 31–34.
[10] K. Swani, B.P. Brown, G.R. Milne, Should tweets differ for {B2B} and B2C? An analysis of fortune 500 companies' twitter communications, Ind. Mark. Manag. 43 (5) (2014) 873–881.
[11] N. Naveed, T. Gottron, J. Kunegis, A.C. Alhadi, Bad news travel fast: a content-based analysis of interestingness on twitter, in: Proceedings of the 3rd International Web Science Conference, WebSci'11, ACM, New York, NY, USA, 2011, pp. 8:1–8:7.
[12] P. Bellot, V. Moriceau, J. Mothe, E. SanJuan, X. Tannier, Overview of INEX tweet contextualization 2013 track, in: Working Notes for CLEF 2013 Conference, Valencia, Spain, September 23–26, 2013, 2013.
[13] P. Bellot, V. Moriceau, J. Mothe, E. SanJuan, X. Tannier, Overview of INEX tweet contextualization 2014 track, in: Working Notes for CLEF 2014 Conference, Sheffield, UK, September 15–18, 2014, (2014), pp. 494–500.
[14] M.-C. Yang, H.-C. Rim, Identifying interesting twitter contents using topical analysis, Expert Syst. Appl. 41 (9) (2014) 4330–4336.
[15] E. Reiter, The Handbook of Computational Linguistics and Natural Language Processing, Blackwell Handbooks in Linguistics, John Wiley & Sons, 2010, pp. 574–598 Ch. Natural Language Generation.
[16] E. Lloret, M. Palomar, COMPENDIUM: a text summarisation tool for generating summaries of multiple purposes, domains, and genres, Nat. Lang. Eng. 19 (2013) 147–186.
[17] T. Vodolazova, E. Lloret, R. Muñoz, M. Palomar, The role of statistical and semantic features in single-document extractive summarization, Artif. Intell. Res. 2 (3) (2013) 35–44.
[18] T. Vodolazova, E. Lloret, R. Muñoz, M. Palomar, Extractive text summarization: can we use the same techniques for any text? in: E. Métais, F. Meziane, M. Saraee, V. Sugumaran, S. Vadera (Eds.), Natural Language Processing and Information Systems, vol. 7934 of Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2013, pp. 164–175.
[19] B. Sharifi, M.-A. Hutton, J. Kalita, Summarizing microblogs automatically, in: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, HLT'10, Stroudsburg, PA, USA, Association for Computational Linguistics, 2010, pp. 685–688.
[20] F. Liu, Y. Liu, F. Weng, Why is SXSW trending?: Exploring multiple text sources for Twitter topic summarization, in: Proceedings of the Workshop on Languages in Social Media, LSM'11, Association for Computational Linguistics, Stroudsburg, PA, USA, 2011, pp. 66–75.
[21] B. Sharifi, M.-A. Hutton, J.K. Kalita, Experiments in microblog summarization, in: Proceedings of the 2010 IEEE Second International Conference on Social Computing, SOCIALCOM'10, IEEE Computer Society, Washington, DC, USA, 2010, pp. 49–56.
[22] D. Chakrabarti, K. Punera, Event summarization using tweets, in: Proceedings of the Fifth International Conference on Weblogs and Social Media, Barcelona, Catalonia, Spain, July 17–21, 2011, 2011.
[23] D. Corney, C. Martin, A. Göker, Two sides to every story: subjective event summarization of sports events using twitter, in: Proceedings of the 1st International Workshop on Social Multimedia and Storytelling Co-located with ACM International Conference on Multimedia Retrieval ICMR 2014, Glasgow, Scotland, UK, April 1, 2014, 2014.
[24] M. Melero, M.R. Costa-Jussà, J. Domingo, M. Marquina, M. Quixal, Holaaa!! writin like u talk is kewl but kinda hard 4 NLP, in: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012), European Language Resources Association (ELRA), 2012.
[25] M. Fuentes, M. Massot, H. Rodríguez, L. Alonso, Mixed approach to headline extraction for DUC 2003, in: Proceedings of Document Understanding Conference (DUC 2003), Edmonton, Alberta, Canada, 2003.
[26] S. Wan, M. Dras, C. Paris, R. Dale, Using thematic information in statistical headline generation, in: Proceedings of the ACL 2003 Workshop on Multilingual Summarization and Question Answering, vol. 12, 2003, pp. 11–20.
[27] B. Dorr, D. Zajic, R. Schwartz, Hedge trimmer: a parse-and-trim approach to headline generation, in: Proceedings of the HLT-NAACL 03 on Text Summarization Workshop, vol. 5, Association for Computational Linguistics, 2003, pp. 1–8.
[28] R. Soricut, D. Marcu, Abstractive headline generation using WIDL-expressions, Inf. Process. Manag. 43 (6) (2007) 1536–1548.
[29] C. Lopez, V. Prince, M. Roche, Just title it! (by an online application), in: Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Stroudsburg, PA, USA, 2012, pp. 31–34.
[30] K. Woodsend, M. Lapata, Automatic generation of story highlights, in: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, 2010, pp. 565–574.
[31] K. Woodsend, Y. Feng, M. Lapata, Title generation with quasi-synchronous grammar, in: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, 2010, pp. 513–523.
[32] P. André, M. Bernstein, K. Luther, Who gives a tweet?: Evaluating microblog content value, in: Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work, CSCW'12, ACM, New York, NY, USA, 2012, pp. 471–474.
[33] C. Hsieh, C. Moghbel, J. Fang, J. Cho, Experts vs the crowd: examining popular news prediction performance on twitter, in: WSDM'13: 6th ACM International Conference on Web Search and Data Mining, Rome, Italy, 2013.
[34] C. Lopez, V. Prince, M. Roche, How can catchy titles be generated without loss of informativeness? Expert Syst. Appl. 41 (4, Part 1) (2014) 1051–1062.
[35] L. Padró, E. Stanilovsky, Freeling 3.0: towards wider multilinguality, in: Proceedings of the Language Resources and Evaluation Conference (LREC 2012), ELRA, Istanbul, Turkey, 2012.
[36] J. Jacoby, M.S. Matell, Three-point Likert scales are good enough, J. Mark. Res. 8 (4) (1971) 495–500, http://dx.doi.org/10.2307/3150242.
[37] S. Petrović, D. Matthews, Unsupervised joke generation from big data, in: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (vol. 2: Short Papers), Association for Computational Linguistics, Sofia, Bulgaria, 2013, pp. 228–232.
[38] S. Mahamood, E. Reiter, Generating affective natural language for parents of neonatal infants, in: Proceedings of the 13th European Workshop on Natural Language Generation, Association for Computational Linguistics, Stroudsburg, PA, USA, 2011, pp. 12–21.
[39] C. Horn, A. Zhila, A.F. Gelbukh, R. Kern, E. Lex, Using factual density to measure informativeness of web documents, in: Proceedings of the 19th Nordic Conference of Computational Linguistics, NODALIDA 2013, May 22–24, 2013, Oslo University, Norway, 2013, pp. 227–238.
[40] E. SanJuan, V. Moriceau, X. Tannier, P. Bellot, J. Mothe, Overview of the INEX 2012 tweet contextualization track, in: in: CLEF 2012 Evaluation Labs and Workshop, Online Working Notes, Rome, Italy, September 17–20, 2012, 2012.
[41] J. Cohen, A coefficient of agreement for nominal scales, Educ. Psychol. Meas. 20 (1) (1960).
[42] A.J. Viera, J.M. Garrett, Understanding interobserver agreement: the kappa statistic, Fam. Med. 37 (5) (2005) 360–363.

[43] K. Gwet, Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring The Extent of Agreement Among Raters, 4th ed., Advanced Analytics, LLC, 2014.
[44] E.F. Tjong Kim Sang, F. De Meulder, Introduction to the CoNLL-2003 shared task: language-independent named entity recognition, in: Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 – vol. 4, CONLL'03, Association for Computational Linguistics, Stroudsburg, PA, USA, 2003, pp. 142–147.
[45] M. Marrero, S. Sánchez-Cuadrado, J.M. Lara, G. Andreadakis, Evaluation of named entity extraction systems, Adv. Comput. Linguist. Res. Comput. Sci. 41 (2009) 47–58.
[46] P.D. Leedy, J.E. Ormrod, Practical Research: Planning and Design, 8th ed., Prentice Hall, 2004.
[47] A.P. Chan, D.W. Chan, Developing a benchmark model for project construction time performance in Hong Kong, Build. Environ. 39 (3) (2004) 339–349.
[48] R. Stockwell, J. Bowen, J. Martin, The Grammatical Structures of English and Spanish, Contrastive Structure Series, University of Chicago Press, 1965.
[49] Verb Phrase Syntax: A Parametric Study of English and Spanish, Studies in Natural Language and Linguistic Theory, Kluwer Acad. Publ., 1988.
[50] M.-J. Cuenca, Two ways to reformulate: a contrastive analysis of reformulation markers, J. Pragmat. 35 (7) (2003) 1069–1093, http://dx.doi.org/10.1016/S0378-2166(03)00004-3.
[51] C. Tan, L. Lee, B. Pang, The effect of wording on message propagation: topic- and author-controlled natural experiments on twitter, in: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (vol. 1: Long Papers), Association for Computational Linguistics, Baltimore, Maryland, 2014, pp. 175–185.
[52] A. Wang, T. Chen, M.-Y. Kan, Re-tweeting from a linguistic perspective, in: Proceedings of the Second Workshop on Language in Social Media, LSM'12, Association for Computational Linguistics, Stroudsburg, PA, USA, 2012, pp. 46–55, http://dl.acm.org/citation.cfm?id=2390374.2390380.
[53] Y. Artzi, P. Pantel, M. Gamon, Predicting responses to microblog posts, in: Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT'12, Association for Computational Linguistics, Stroudsburg, PA, USA, 2012, pp. 602–606, http://dl.acm.org/citation.cfm?id=2382029.2382126.

Dr. Elena Lloret is an assistant full-time lecturer at the University of Alicante in the area of databases. She obtained her Computer Science graduate and PhD at the University of Alicante. Her main field of interest is text summarization, text simplification and natural language generation. She is the author of over 30 scientific publications in relevant journals and international conferences. She has been collaborating with international researchers and has participated in a number of projects at a national level (ATTOS (TIN2012-38536-C03-03), LEGOLANG-UAGE (TIN2012-31224)), as well as European project, such as FIRST (grant no. FP7-287607) or SAM (FP7-611312). She has also collaborated with international groups in Madrid, Wolverhampton, Sheffield, Edinburgh, and Texas.

Prof. Dr. Manuel Palomar is the University President of the University of Alicante and head of the Natural Language Processing and Information Systems Research Group of the same university. He is also a full professor of this University since 1991 and his main teaching area focuses on the analysis, design and management of databases, datawarehouses, and information systems. He received his Master's degree and Ph.D. in Computer Science at the Polytechnic University of Valencia, Spain. His research interests are Human Language Technologies (HLT) and Natural Language Processing (NLP), in particular Text Summarization, Semantic Roles, Textual Entailment, Information Extraction and Anaphora Resolution. He has supervised more than 12 thesis and he is the author of more than 70 scientific publications on international journals and conferences on different topics related to HLT and NLP. Furthermore, he has coordinated and been involved in a number of regional, national and international research projects funded by the Generalitat Valenciana (Valencian Government), the Ministry of Science and Innovation (Spanish Government) and the European Council.