A Computational Exploration of Exaggeration

Enrica Troiano and **Carlo Strapparava** and **Gözde Özbal** and **Serra Sinem Tekiroğlu**
FBK-Irst, Trento, Italy
enrica.troiano@gmail.com, strappa@fbk.eu, gozbalde@gmail.com, tekiroglu@fbk.eu

Abstract

Several NLP studies address the problem of figurative language, but among non-literal phenomena, they have neglected exaggeration. This paper presents a first computational approach to this figure of speech. We explore the possibility to automatically detect exaggerated sentences. First, we introduce HYPO, a corpus containing overstatements (or hyperboles) collected on the web and validated via crowdsourcing. Then, we evaluate a number of models trained on HYPO, and bring evidence that the task of hyperbole identification can be successfully performed based on a small set of semantic features.

1 Introduction

Very often, when humans recount their experiences, describe objects or verbalize ideas, they exaggerate in some respects. "*All you need is love*", sang The Beatles in one of their most iconic recordings; many centuries before, Shakespeare had Romeo say that there was "*more danger in Juliet's eyes than in twenty swords*"; and Maximus Meridius, a Roman commander in the movie The Gladiator, incited his legion to battle by the words "*At my signal, unleash hell*". Exaggerating, like in these quotes, is a linguistic tendency that unfolds in a variety of situations. From TV advertisements to debates in politics, our verbal productions are infused with statements, or more precisely overstatements, that puff up facts.

The study of exaggeration dates back to ancient Greece, and its centrality in the spectrum of figures of speech has long been established. Empirical evidence has shown that it is the most used rhetorical device, second only to metaphor (Kreuz et al., 1996), but in comparison to its kin tropes, this phenomenon represents an under-researched field in NLP. In effect, the problem of automatic detection of exaggerations (or hyperboles) has been quite dismissed. On the one hand, determining that a sentence speaks an excess is often a complex and context dependent act, on the other, no resource has ever been made available to specifically address the figure. To overcome this issue, we build HYPO, the first dataset focused on exaggeration, and introduce the task of automatic hyperbole detection. Specifically, the goal of this work is to find an automatic solution to establish if the topic of a sentence is aggrandized or, on the contrary, is presented as it is in reality.

Effective identification of overstatements would benefit theoretical and applied approaches to natural language. It can represent both a deterministic strategy to test hypotheses about exaggerations, as well as a boost for dialogue systems, information extraction, and all such AI endeavors to understand how humans talk. In fact, automating the intuition that a text says more than is true can meet a number of useful applications. Endowed with this ability, automatic tools might support our recognition of some kind of fake news, which blow information out of proportion; they might ascertain whether the promises of testimonials and politicians constitute a form of puffery; they might even be exploited for health-related objectives, as auxiliary tools for the diagnosis of psychological conditions (e.g., depression, narcissistic disorder) which use exaggerations to depict extremely unnuanced views of the world.

To tackle the problem, the paper starts with a synthesis of past theoretical work, which provides a guideline for building HYPO. Data was collected by a manual crawl on the web and its quality was tested via crowdsourcing. We then use the overstatements of HYPO for learning to classify a sentence as hyperbolic or literal. Experimental results show that a number of models can solve the problem with above chance accuracy, thus confirming that the task is feasible and that HYPO can be ex-

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 3296–3304
Brussels, Belgium, October 31 - November 4, 2018.

ploited for similar purposes in future work.

2 Related Work

A long lasting challenge in NLP is to reason about figurative language, and in fact, scientists have achieved groundbreaking results in the study of a range of figures. Nevertheless, they have given scarce attention to hyperbole – if any at all. Only relatively recently has this trope become an object of interest of corpus linguistics, and some insights on its nature have been delivered by statistics extracted from written and spoken sources.

The bulk of research points out that a hyperbole is not perceived nor produced as a lie. It makes things sound bigger than they actually are, but with the goal of emphasizing them, of clarifying ideas, of adding interest and humor to the conversation (Roberts and Kreuz, 1994). In other words, speakers exaggerate to accentuate the one element, account or aspect of reality that is important to them, in order to make others see truth from their perspective (Ritter, 2012). Hearers, on their part, suspend disbelief towards the literal content of the hyperbolic expression, which is mendacious, and recover its intended meaning, which is instead in accord with reality. Basically, to find this non-literal content, they reformulate a hyperbole as a paraphrase devoid of figures of speech, which is softer than the hyperbole itself and still conveys a strong concept (Fogelin, 1988).

Another point on which the literature agrees is that hyperboles are pragmatic acts. They may entirely rely on the concrete context of their production, like in the statement "It took ages to build the castle", which carries both an exaggerated and a literal sense, depending on whether it refers to a playful sand castle or to a strong walled construction. McCarthy and Carter (2004) endorse this view based on the need of contextual, extra-linguistic information, and claim that understanding an overstatement is to perceive a 'contrast' between the expression itself and its referent, i.e. a discrepancy between reality as it is and as is described.

Focusing on this contrast, Cano Mora (2010) provides a handcrafted framework of the semantic fields in which it tends to arise. Her taxonomy develops along a quantitative and a qualitative dimensions, suggesting that any exaggeration inflates either a measurable or a subjective property of the topic of discourse. However, given that any type of speaking non-literal is a departure from actual facts, the notion of 'contrast' can be referred to figures of speech in general. With this regard, an important point is made about the peculiarities of hyperbole (Colston and O'Brien, 2000; Carston and Wearing, 2011). Exaggerating presents facts with a greater degree, e.g. being bigger, more desirable, so it prompts a contrast of magnitude. Other tropes engender a contrast of kind for they portray an object via some types of qualities that it does not actually have. Metaphors, for instance, borrow those qualities from the vehicle term.

Therefore, the major contribution yielded by past corpus research is to have made explicit a few typically hyperbolic characteristics, which will come handy for the construction of our corpus.

3 HYPO

Hyperboles over-blow the truth, by augmenting or down-toning the qualities of the referent of discourse, e.g., an event, an object, a person, etc. If a referent has a feature X, a hyperbole presents it as having more of that X than warranted by reality (Claridge, 2011): what is big becomes bigger (e.g. [1]) and what is small becomes smaller (e.g. [2]). This causes a discrepancy between the linguistic expression and the actual state of affairs, and provides the former with a non-literal meaning.

[1] Her morning jog turned into a marathon.

[2] I'm ready in no time.

Drawing on example from (McCarthy and Carter, 2004), we use this 'contrast' or 'counterfactuality' as the central condition that makes a statement hyperbolic. It can be noticed that the extent of the contrast corresponds to different types of overstatements, which can go from slight distortions of real situations (e.g. [1]), to representations of absurd worlds (e.g. [2]). In the second instance, the contrast is grasped with no need for pragmatic information, but other types of overstatements can be perceived as such only if one knows the contextual setting in which they were produced. In those cases, readers evaluate the credibility of the sentence, that is, they confront how the sentence presents property X of the topic to how they expect it to be, given their previous experience and their common knowledge about the world (Ferré, 2014). For instance, although one cannot verify if proposition [1] is in contrast with the truth, hardly would one believe that a routine jog has equaled a marathon.

[3] I won't wait for you: it took you centuries to get dressed.

[4] That bag is to die for!

The above examples picture some peculiar traits of hyperboles. First, they require the possibility to present concepts with different intensities, by shifting ideally towards the extremes of a continuous semantic scale. Second, there are two scales on which this shift occurs (Cano Mora, 2010). One is quantitative, and it allows to inflate objective, measurable features (e.g. time [3]), while the other is qualitative and serves to augment subjective characteristics (e.g. appreciation [4]).

Another defining element of hyperboles is their emotional bent. When we exaggerate, we accentuate some pieces of evidence that supports our thoughts and perspectives, while downplaying competing alternatives. We present an extravagant view of reality with the goal of manifesting our positive or negative involvement towards it. This might appear obvious for qualitative hyperboles such as [4], for the property they enlarge is subjective, but quantitative overstatements have an evaluative trait as well. As an example, the overstatement [3] frames an antipathetic position by increasing an impartial, unemotional measure.

To sum up, there are three criteria that were deemed to characterize a hyperbole. They are the non-literal meaning, the upsurge on a semantic scale and a connotative trait. Hyperbolic candidates entered the corpus only if they were considered to have a figurative content rather than a deceptive one, if the figurative component emerged as a contrast of magnitude with reality, and if the contrast seemed to color reality with an evaluative tone.

3.1 Dataset Description

HYPO is a collection of 709 objects that comply with the three definitional requirements, and that can be deemed exaggerated without knowing the context in which they were originally produced[1].

Sentences can be semantically grouped into basic and composite items, which respectively preserve and modify the semantic domain of their referent (Claridge, 2011). For instance, [5] is basic, because the intended and literal meanings have the same domain (i.e. size) though they differ in magnitude. Instead, the exclamation in [6] is composite, as the intended meaning concerns human emotions but is expressed with a quality inherent to stones. This results both in a metaphor, since it describes a psychological state by means of a domain transfer, and in a hyperbole, because among all the things that have limited movement capabilities, it pinpoints the stones, which are completely motionless.

[5] The house is the size of a postage stamp.

[6] First I was afraid, I was petrified!

[7] I avoid crowded places like the plague.

[8] She agreed with every word of my argument.

Other hyperboles arise from the combination with figures different than metaphors, such as comparatives [7] and synecdoches [8].

From a syntactic point of view, exaggerations in the data either spread over phrases, or pop up as single words that belong to any sorts of grammatical classes. One peculiar instance is that of Extreme Case Formulations (ECFs), which are adjectives ("absolute, whole"), quantifiers and nouns ("all, no, everybody"), adverbs ("always, never"), phrases ("as good as it gets") and superlative constructions ("ever, the most") that engender extreme utterances by evoking the highest degree on a semantic scale (Pomerantz, 1986). Although ECFs might not be "heard as absurd or counterfactual and often display a degree of conventionality" (McCarthy and Carter, 2004), they share the three fundamental characteristics of the trope: they build expressions around semantic acmes, which lessens their credibility and prompts hearers to grasp their non-literal content, namely, one that expresses an evaluation. Some ECFs are at work in the following propositions.

[9] This is the best pizza in history.

[10] Everybody loves chocolate.

Within the above groups lie qualitative [9] and quantitative [5] hyperboles, creative [7] and conventional [10] ways of exaggerating. Notably,

[1] As an example, rather than "It took ages to build the castle", which can be read as hyperbolic or literal depending on contextual information, our dataset would be appended with "It took ages to build the castle. After a few minutes, my little brother had already destroyed it!", which is non-literal regardless of the context of its production.

such categories can criss-cross (Peña and Ruiz de Mendoza, 2017) like in "I'd bake cakes again and again and again", an instance of a quantitative hyperbole that encompasses multiple tokens, as it is merged with a polysyndeton.

4 Resource Construction

The collection of data for HYPO proceeded along two lines. One involved our own effort to invent overstatements. The other consisted of a manual crawl of the Internet and targeted the scripts of animated cartoons, advertisements, love letters, clickbait headlines, as well as other material that allegedly fulfilled the communicative objectives of exaggerations (see Roberts and Kreuz, 1994), and therefore was likely to include instances of this figure of speech. A total of 804 overstatements was obtained which comply with the definition presented in Section 3.

These candidate items underwent a validation stage on the Crowdflower/Figure Eight platform[2], where their hyperbolicity was judged by external annotators, and the final corpus was determined. Ideally, our microtask was feasible by any English speaker. It asked workers to read 10 sentences, one of which was a test item, and to answer 6 questions aimed at determining if and how the texts overstate their topic. Gold units consisted of 42 hyperboles, together with 50 non-hyperbolic statements to ensure that raters would not learn to mark all items as exaggerated. Each of the 854 sentences was expected to be judged by 5 annotators provided with minimal instructions on the subject at hand.

In particular, the annotators decided if a sentence contained a hyperbole (*Question 1*), and if so, they highlighted the words that caused its hyperbolic bent (*Question 2*), paraphrased it without exaggerating (*Question 3*), classified the hyperbole as quantitative or qualitative (*Question 4*), rated the degree of the exaggeration (*Question 5*) and established whether it was creative or conventional (*Question 6*).

With the first question, candidate items for HYPO were validated. In the second, annotators selected the smallest number of words (or minimal units) that convey the exaggeration, like "million" in "I've told you a *million* times". This assignment enabled us to collect the units perceived as hyperbolic, and to use them for creating another corpus, that is, a dataset of sentences which include the hyperbolic tokens of HYPO, but without their hyperbolic twist (see 4.2). Similarly, the rationale beyond the third question was the construction of a corpus devoid of exaggerations. People had to paraphrase the sentences without exaggerating. They were encouraged to reword or delete the portion of text highlighted in *Question 2*, such that the resulting sentences would have differed from the original ones only for the absence of hyperbolic tokens. Basically, we explored the idea that understanding a hyperbole requires a sort of readjustment, where concepts are toned down, but still to a high degree. For instance, an overstatement such as "I've told you a *million* times" could have been modified as in "I've told you a lot of times".

In *Question 4*, the annotators decided if a hyperbole belonged to a quantitative or a qualitative class, according to whether it exaggerated something which is objective and quantifiable with a number (e.g. "He *never* says no") or which is subjective and unmeasurable (e.g. "He seems to *come from another world*!"). This was meant to label sentences with one of the two semantic dimensions along which a concept is aggrandized, as proposed in the literature (Cano Mora, 2010). As for *Question 5*, an exaggeration was rated as either 'Possible', if it denoted an extreme but conceivable situation (e.g. "I avoid crowded places like the plague"), or 'Impossible', when it described an absurd or paradoxical situation (e.g. "My father always works"). We collected these scores to test the hypothesis that sentences with a higher hyperbolic degree are classified more accurately in the experiment because their excess is easier to detect. In the sixth question, exaggerations were judged either as conventional ("He died of envy") or creative ways to express an idea ("It got so cold that all spoken words froze solid"). The answers were expected to show if there is a correlation between conventional and 'Possible' hyperboles of *Question 5* (i.e. if conventional items are perceived as softer exaggerations)[3].

4.1 Evaluation Measures

Only 750 sentence received 5 reliable judgments. Therefore, to evaluate the quality of the results, the

[2] https://www.figure-eight.com

[3] For the quality check, we only required the raters to correctly judge questions 1 and 2. Question 3 was not included as no unique answer exists. Questions 4, 5 and 6 were also not included because, due to their difficulty, we expected some amount of disagreement among the raters.

inter-annotator agreement was observed on those. We calculated two measures of agreement for the task of recognizing hyperboles (Question 1). First, we measured the raw agreement (RA) that is the proportion of items with unanimous judgments (i.e. the sentences marked as either hyperbolic or non-hyperbolic by all the workers) out of the total number of items with 5 judgments. The final score showed that annotators agreed in 58.5% of the cases, which we considered an acceptable metric for the reliability of the annotation process.

That people had the same understanding of hyperboles in more than half the items seemed reasonable, because it can be challenging to establish if a statement is extreme, especially for non-expert workers. In fact, contextual knowledge, which in everyday situations enables interlocutors to grasp the inflation of a sentence, is neglected in an experiment that revolves around isolated textual passages. Moreover, the RA measure is extremely sensitive to the inconsistency of results, as the presence of one incongruous annotation out of 5 is enough to lower the outcome.

Therefore, we used a second measure to take into consideration the difficulty of having 5 people agree. Since items were labeled by more than two annotators, we computed a pairwise agreement, as suggested by (Artstein and Poesio, 2008). We exploited the Observed Agreement (A_O) as defined in Fleiss's κ, which consists of the proportion of items on which pairs of annotators agree out of the total number of judgment pairs. Also in this case we observed a substantial agreement (A_O=80.2%) corroborating the quality of our results.

4.2 Final Datasets

The logic behind the validation of our resource was to obtain adequate information to train a classifier of hyperboles. Given the consistency of annotations, we selected the sentences to be inserted into the corpus of hyperboles and built two control datasets. The outcome is a set of hyperboles (HYPO) with two types of non-figurative counterparts, that is, their paraphrases, and literal sentences that include the words that are hyperbolic in HYPO, but with a literal connotation. By means of illustration, while HYPO would incorporate a sentence like "Her morning jog turned into a marathon", the Paraphrases corpus and the Minimal Units Corpus would respectively include "Her morning jog turned into a very long run" and "There is a marathon in the city today".

HYPO The sentences in HYPO are the items that at least 3 annotators out of 5 perceived as exaggerations. From the total of 854 judged sentences, we discarded the non-hyperbolic units, those that received less than 5 judgments and those which were not classified as hyperbolic by the majority of annotators. The final result is a corpus of 709 hyperboles[4].

Paraphrases Corpus The corpus of paraphrases was created by choosing one paraphrase for each hyperbolic sentence, i.e. the one that introduced the smallest change in the syntax and the semantics of the corresponding hyperbole. The result is a collection of 709 sentences which say the same things as their hyperbolic counterparts but are devoid of exaggerations.

Minimal Units Corpus To build this dataset, we used the minimal units selected by the annotators in Question 2. We wanted to end up with sentences which are not hyperbolic and which contain the same hyperbolic words of HYPO.

For each exaggeration we considered the tokens that were selected by the majority of annotators. They either consisted of a single term, or a phrase, or long-distance words. Then, we extracted sentences containing these tokens from sources such as the WaCKy corpus, a dump of the English Wikipedia, whose editorial criteria force the entries to be neutral and verifiable, and thus, unlikely to incorporate excessive statements. Whenever this approach did not produce results, we ran a Google search. The final corpus contains 698 non-hyperbolic sentences.

5 Experiment

The task of hyperbole detection is formulated as a supervised learning problem, and specifically, as a sentence-based classification with two classes, that is, hyperbolic and literal.

5.1 Features Set

To describe sentences, we define two sets of features that incorporate the qualitative and quantitative criteria proposed in the literature[5].

[4] The dataset will be made publicly available under a Creative Commons License for free cultural works.

[5] These terms do not refer to the semantic scale on which the shift of meaning occurs, which vary depending on the topic of hyperboles, but to the counterfactuality and the connotative traits that characterize them all.

All of the features take on values in the interval [-1,1], and some are encoded in multiple ways to ensure richer information. As for the quantity group, the notion that hyperboles say *more of X* when *X* is the case is decomposed into two features, i.e. imageability and unexpectedness, while the qualitative marker of hyperboles, or the view that they shape a speaker's perspective about *X*, is rendered by the polarity, the subjectivity and the emotional strength of sentences.

Imageability is the degree to which a word can evoke a mental image. Speakers hyperbolize to convey meanings with strength, and we assumed that such a goal might be backed by a highly picturable vocabulary. This feature is extracted from the resource of Tsvetkov et al. (2014), who propagated the imageability ratings of the MRC psycholinguistic database to 150.114 terms. For each sentence, we averaged the imageability values of all its words.

Unexpectedness refers to the fact that hyperboles are less predictable expressions than literals. Basically, minimal hyperbolic units modify the real characteristics of *X*, and in this sense, they are incoherent with the rest of discourse about *X*: they are out of context, and come unexpected to the hearers of overstatements.

We conjecture that word vectors may capture if an expression is being used "unexpectedly" because they encode the contexts in which terms frequently occur, as well as contrasts and similarities among their meanings. In fact, according to the Distributional Hypothesis (Harris, 1954), lexical items appearing in similar contexts tend to have similar meanings. Our expectation is that the words of a figurative sentence carry less similar meanings compared to those of a literal instance, and hence, that their vectorial representations turn out to be more distant.

For every sentence in the dataset, we map its words onto both the pre-trained vectors of Mikolov et al. (2013), obtained from the Skip-gram model, and the GloVe vectors by Pennington et al. (2014). Then, we consider the cosine distance between all possible word pairs to score their semantic similarity. The unexpectedness feature of a sentence is found in two ways: as the average similarity among all of its word pairs, and as the lowest of those pair similarities. Both measures are separately computed with Skip-Gram and GloVe vectors, resulting in 4 scores.

Polarity corresponds to the sentiment of a statement. It is extracted through both TextBlob (Loria, 2014), a system that directly scores sentences, and SentiWords (Gatti et al., 2016), which lists polarity values for 155k POS-tagged lemmas. With this lexicon, we find the sentiment of a sentence by averaging the sentiment of all of its word lemmas.

Subjectivity, found with TextBlob, which specifies if a statement conveys an objective information or a personal opinion. Ideally, the higher the absolute value of polarity in the range [0,1], the more subjective a sentence.

Emotional intensity stands for the strength of sentiment. It captures if the utterer of a sentence is sympathetic towards the thing being said, while quantifying the emphasis with which such position is communicated. The scores are obtained from VADER (Hutto and Gilbert, 2014).

5.2 Experimental Setup and Results

A simple pre-processing is applied to HYPO, the Paraphrases and the Minimal Units corpora prior to the experiment. We remove the stopwords, for which the quality scores are not available. The resulting sentences are represented by 9-dimensional vectors, in which 5 entries stand for the quantity features and 4 belong to the quality group.

For classification, we experiment with various algorithms, such as Logistic Regression (LR), Naive Baies (NB), k-Nearest Neighbors (KNN), Decision Trees (DT), Support Vector Machine (SVM) and Linear Discriminant Analysis (LDA). The models are evaluated by comparing their average accuracy on a 10-fold cross-validation against three baselines: a classifier that randomly associates inputs to the hyperbolic and literal labels, one that classifies sentences using the 300 features of the pre-trained Skip-gram representations, and a third that relies on the 300-dimensional GloVe vectors trained by their authors. Sentences fed to the baseline classifiers were represented as the sum of their word vectors[6].

[6] Other types of compositional methods yield results analogous to those in Table 1 and Table 2.

	Baseline1	QQ	Skip-gram	GloVe	Skip-gram+QQ	GloVe+QQ
LR	.50	.64	.68	.66	**.72**	.69
KNN	.50	.63	.47	.43	.52	.48
NB	.50	.66	.69	.66	.69	.68
DT	.50	.60	.54	.53	.55	.54
SVM	.50	.64	.15	.62	.63	.64
LDA	.50	.61	.67	.65	.68	.67

Table 1: Mean accuracy of 10-fold cross validation in the Hype-Par setting. Column QQ shows results for our handcrafted quality and quantity features; the last two columns are concatenations of QQ with the Skip-gram and GloVe baseline features.

	Baseline1	QQ	Skip-gram	GloVe	Skip-gram+QQ	GloVe+QQ
LR	.50	.44	.60	.58	.61	.59
KNN	.50	.47	.49	.48	.50	.51
NB	.50	.50	.64	.64	.62	**.68**
DT	.50	.51	.51	.52	.56	.54
SVM	.50	.02	.09	.59	.13	.60
LDA	.50	.52	.54	.34	.57	.56

Table 2: Mean accuracy of 10-fold cross validation in the Hype-Min setting.

Cross-validation is conducted in two settings. In one (Hype-Par), the non-hyperbolic sentences are paraphrases, in the other (Hype-Min), literal data come from the Minimal Units Corpus. The results are shown in Table 1 and Table 2.

While in the Hype-Min setting the performance is not satisfying, estimators achieve above chance accuracy using paraphrases as literal inputs. In fact, in Table 1, the accuracy scores based on quantity-quality vectors (QQ column) suggest that our handcrafted features are actually useful for detecting hyperboles. Therefore, to gain further insight on their informativeness, we conduct a recurrent feature ablation and observed how different subsets affect predictions. Figure 1 illustrates that 5 features can maximize the accuracy of LR. SVM and LDA behave the same with a set of the same size, and they all assign high weights to imageability, unexpectedness and subjectivity. The three models become comparable to, and yet do not outperform, the second and third baselines.

In the attempt to improve the models' description of the data, we repeat the experiment with yet another set of features. We merge the QQ with the Skip-Gram and GloVe features, by separately concatenating the two types of vectors to our data representations (Skip-Gram+QQ and GloVe+QQ columns).

An interesting trend appears both for Hype-Par and Hype-Min: with Skip-Gram+QQ, algorithms perform better than relying on Skip-Gram or QQ alone, and the same happens for Glove+QQ. The new sets of features produce a consistent improvement over the baselines and over our own features. LR outstands other classifiers in the Skip-Gram+QQ combination, reaching .72 mean accuracy and .76 average F1-score (see Table 3).

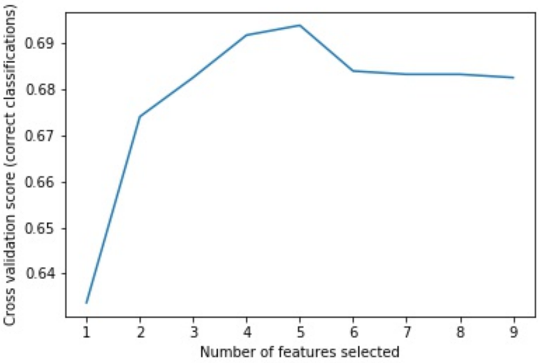

Figure 1: Recurrent Feature Elimination with LR in the Hype-Par setting.

5.3 Analysis

The concatenation of vectors enhance performances and it provides evidence that our quantity and quality vectors enrich both Skip-Gram and GloVe with some useful information about hyperbole. This observation, together with the outcome of LR, suggests that the task introduced in the

	Hype-Par (LR, Skip-Gram+QQ)	Hype-Min (NB, GloVe+QQ)
Precision	.76	.54
Recall	.76	.53
F1-score	**.76**	.52

Table 3: Average precision, recall and F-1 values obtained in cross validation for the best performing models in Hype-Par and Hype-Min.

	Conventional	Measurable	Possible
YES	50	18	20
NO	43	75	73

Table 4: Annotators' judgments about 93 misclassified hyperboles.

present work delineates a promising field of inquiry, and that it can benefit from our QQ features.

To deepen our understanding of the results, we analyze the errors made in the Hype-Par classification. This conclusive stage of the study serves to probe if the overstatements that tend to be misclassified share some characteristics that are missing from our group of features.

From a test set of 468 data points, we collect those that are incorrectly labeled by all of the models, which comprise 185 sentences, 93 of which are hyperbolic. Examples are: "He's more aged than the hills" (Hype), "They will die of envy" (Hype), "You get into that university, you won't get out alive" (Hype), "He hiccuped for a long time" (Par).

We investigate the characteristics for which we collected judgments in the Crowdflower experiment (Questions 4, 5 and 6 in the Section above), that is, the measurable trait of exaggerations, and their degree of hyperbolicity and conventionality. Table 4 details the judgments about misclassified hyperboles, as rated by the the majority of their annotators. More than half hyperboles were declared conventional, impossible and non-measurable.

It appears that conventional hyperboles are more difficult to recognize (the test set comprises an equal number of conventional and creative items). This is not surprising if we consider that the classifiers have especially relied on the unexpectedness feature. Vectors encode information relative to the context where words occur, so they might capture that the words of a conventional hyperbole are likely to be used together, since they are highly common in language (e.g. [2]).

As for the second characteristic, the majority of hyperboles amplifies, according to the annotators, an unmeasurable trait of the topic of discourse. This sheds light on the topic of the errors: understanding that a sentence overshoots reality with respect to a subjective impression (e.g. heaviness of commitment [3]) is harder than with an objective quality (e.g. age [1]). In future work, we may investigate how to better formalize the counterfactuality condition, by specifying different strategies to use in the two cases.

Lastly, errors regarding the degree of hyperbolicity run up against our expectations. We hypothesized that a more exaggerated hyperbole (i.e. impossible) is easier to identify, but statistics suggest the opposite. Impossible hyperboles may be obvious for humans, who use pragmatic knowledge, but not for an agent which entirely relies on linguistic information. Prospective research may test if this problem can be overcome with the help of multimodal strategies.

6 Conclusions

Hyperbole, the figure of exaggeration and one of the hallmarks of human communication, is tackled in this paper from a computational perspective. Our research aimed at answering the question whether it is possible to endow a system with the ability to identify exaggerated sentences. Experimental results showed promising directions for their automatic detection and suggest that the execution of this task can be based on semantic features. As a novel approach to hyperboles, the project started with no related studies to compare to, nor useful resources to investigate. Hence, its main contribution to the field of NLP is the proposal of a new task together with the construction of a corpus of hyperboles, and its main achievement is the devising of a procedure to learn such figurative mechanism. Specifically, the described experiment tested the hypothesis that quantity and quality, which emerge from the body of literature as two core features of exaggerations, are also useful for the automatic processing of the figure.

As a future work, we plan to investigate if our QQ-based models are enhanced by extra-linguistic knowledge, and to incorporate contextual, multimodal features along semantic ones.

References

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Laura Cano Mora. 2010. All or nothing: A semantic analysis of hyperbole. *Revista de Lingüística y Lenguas Aplicadas*, 4(1):25–35.

Robyn Carston and Catherine Wearing. 2011. Metaphor, hyperbole and simile: A pragmatic approach. *Language and Cognition*, 3(2):283–312.

Claudia Claridge. 2011. *Hyperbole in English: A Corpus-Based Study of Exaggeration*. Cambridge University Press.

Herbert L. Colston and Jennifer O'Brien. 2000. Contrast of kind versus contrast of magnitude: The pragmatic accomplishments of irony and hyperbole. *Discourse processes*, 30(2):179–199.

Gaëlle Ferré. 2014. Multimodal hyperbole. *Multimodal Communication*, 3(1):25–50.

Robert J. Fogelin. 1988. *Figuratively Speaking: Revised Edition*. New Haven: Yale University Press.

Lorenzo Gatti, Marco Guerini, and Marco Turchi. 2016. Sentiwords: Deriving a high precision and high coverage lexicon for sentiment analysis. *IEEE Transactions on Affective Computing*, 7(4):409–421.

Zellig S. Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.

Clayton J. Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International Conference on Weblogs and Social Media (ICWSM-14)*.

Roger J. Kreuz, Richard M. Roberts, Brenda K. Johnson, and Eugenie L. Bertus. 1996. Figurative language occurrence and co-occurrence in contemporary literature. In Roger J. Kreuz & Mary Sue MacNealy, editor, *Empirical Approaches to Literature and Aesthetics*, pages 83–97. Norwood, NJ: Ablex Publishing Corporation.

Steven Loria. 2014. Textblob: simplified text processing. Retrieved 04 March 2017 from `http://textblob.readthedocs.io/en/dev/`.

Michael McCarthy and Ronald Carter. 2004. "There's millions of them": hyperbole in everyday conversation. *Journal of pragmatics*, 36(2):149–184.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *ICLR Workshop*.

M^{a} Sandra Peña and Francisco J. Ruiz de Mendoza. 2017. Construing and constructing hyperbole. *Studies in Figurative Thought and Language*, 56:41.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Anita Pomerantz. 1986. Extreme case formulations: A way of legitimizing claims. *Human studies*, 9(2):219–229.

Joshua R. Ritter. 2012. Recovering hyperbole: Rethinking the limits of rhetoric for an age of excess. *Philosophy & Rhetoric*, 45(4):406–428.

Richard M. Roberts and Roger J. Kreuz. 1994. Why do people use figurative language? *Psychological science*, 5(3):159–163.

Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. 2014. Metaphor detection with cross-lingual model transfer. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 248–258.