Detecting Sarcasm in Multimodal Social Platforms

Rossano Schifanella
University of Turin
Corso Svizzera 185
10149, Turin, Italy
schifane@di.unito.it

Paloma de Juan
Yahoo
229 West 43rd Street
New York, NY 10036
pdjuan@yahoo-inc.com

Joel Tetreault
Yahoo
229 West 43rd Street
New York, NY 10036
tetreaul@gmail.com

Liangliang Cao
Yahoo
229 West 43rd Street
New York, NY 10036
liangliang@yahoo-inc.com

arXiv:submit/1633907 [cs.CV] 8 Aug 2016

ABSTRACT

Sarcasm is a peculiar form of sentiment expression, where the surface sentiment differs from the implied sentiment. The detection of sarcasm in social media platforms has been applied in the past mainly to textual utterances where lexical indicators (such as interjections and intensifiers), linguistic markers, and contextual information (such as user profiles, or past conversations) were used to detect the sarcastic tone. However, modern social media platforms allow to create multimodal messages where audiovisual content is integrated with the text, making the analysis of a mode in isolation partial. In our work, we first study the relationship between the textual and visual aspects in multimodal posts from three major social media platforms, i.e., Instagram, Tumblr and Twitter, and we run a crowdsourcing task to quantify the extent to which images are perceived as necessary by human annotators. Moreover, we propose two different computational frameworks to detect sarcasm that integrate the textual and visual modalities. The first approach exploits visual semantics trained on an external dataset, and concatenates the semantics features with state-of-the-art textual features. The second method adapts a visual neural network initialized with parameters trained on ImageNet to multimodal sarcastic posts. Results show the positive effect of combining modalities for the detection of sarcasm across platforms and methods.

Keywords

Sarcasm; Social Media; Multimodal; Deep Learning; NLP

1. INTRODUCTION

Sarcasm is a peculiar form of sentiment expression where the surface sentiment differs from the implied sentiment. Merriam-Webster[1] defines sarcasm as *"the use of words that mean the opposite of what you really want to say especially in order to insult someone, to show irritation, or to be funny."* Sarcasm is a common phenomenon in social media platforms, and the automatic detection of the implied meaning of a post is a crucial task for a wide range of applications where it is important to assess the speaker's real opinion, e.g., product reviews, forums, or sentiment analysis tools.

[1]http://www.merriam-webster.com/dictionary/sarcasm

Figure 1: Example of an Instagram post where the image is needed to detect the sarcasm. The observation "rubbish weather" can only been interpreted correctly by *looking* at the picture. Same holds for "rubbish city."

Most approaches to sarcasm detection to date have treated the task primarily as a text categorization problem, relying on the insight that sarcastic utterances often contain lexical indicators (such as interjections and intensifiers) and other linguistic markers (such as nonveridicality and hyperbole) that signal the sarcasm. In modern online platforms, hashtags and emojis are common mechanisms to reveal the speaker's true sentiment. These purely text-based approaches have been shown to be fairly accurate across different domains [6, 13, 30, 9, 29].

However, in many occasions this text-only approach fails when contextual knowledge is needed to decode the sarcastic tone. For example, in Figure 1, "rubbish weather" is the opposite of what the image represents (i.e., beautiful weather). Without this image, the text could be interpreted as a negative comment about the weather in Liverpool. Recently, several approaches [2, 27, 17, 19, 37] have integrated contextual cues (e.g., the author's profile, author's past posts

and conversations) with the in-post text, showing consistent improvements when detecting sarcasm.

Previous approaches have failed to consider the media linked to the posts as a possible source of contextual information. Tweets, for example, can have audiovisual content attached to the text. Multimodality is the combination of modes of communication (i.e., text, images, animations, sounds, etc.) with the purpose to deliver a message to a particular audience, and it is present in all major social media platforms.

In this work, we leverage the contextual information carried by visuals to decode the sarcastic tone of multimodal posts. Specifically, we consider two types of visual features with different model fusion methods for sarcasm detection. The first approach exploits visual semantics trained on an external dataset, and concatenates the semantics features with state-of-the-art text features. The second method adapts a visual neural network initialized with parameters trained on ImageNet to multimodal (text+image) sarcastic posts. In both methods, we find that visual features boost the performance of the textual models.

We summarize our main contributions as follows:

- We study the interplay between textual and visual content in sarcastic multimodal posts for three main social media platforms, i.e., Instagram, Tumblr and Twitter, and discuss a categorization of the role of images in sarcastic posts.
- We quantitatively show the contribution of visuals in detecting sarcasm through human labeling. This data will be shared with the research community.
- We are the first to propose and empirically evaluate two alternative frameworks for sarcasm detection that use both textual and visual features. We show an improvement in performance over textual baselines across platforms and methods.

We first discuss related work in Section 2. We then describe our data in Section 3, and introduce a categorization of the different roles images can play in a sarcastic post in Section 4. In the same section, we describe how we collect human judgments to build a gold set, and analyze the distribution of posts with respect to the proposed categories. Section 5 describes the details of the two methods for sarcasm detection, and Section 6 presents the experiments carried out to evaluate the frameworks, and their results. Finally, Section 7 concludes the paper, and points to future work.

2. RELATED WORK

Sarcasm as linguistic phenomenon. While the use of irony and sarcasm is well studied from its linguistic and psychological aspects [12], automatic recognition of sarcasm has become a widely researched subject in recent years due to its practical implications in social media platforms. Starting from foundational work by Tepperman et al. [32] which uses prosodic, spectral (average pitch, pitch slope), and contextual (laughter or response to questions) cues to automatically detect sarcasm in a spoken dialogue, initial approaches mainly addressed linguistic and sentiment features to classify sarcastic utterances. Davidov et al. [6] proposed a semi-supervised approach to classify tweets and Amazon products reviews with the use of syntactic and pattern-based features. Tsur et al. [34] focus on product reviews and try to identify sarcastic sentences looking at the patterns of high-frequency and content words. González-Ibáñez et al. [13] study the role of lexical (unigrams and dictionary-based) and pragmatic features such as the presence of positive and negative emoticons and the presence of replies in tweets. Riloff et al. [30] present a bootstrapping algorithm that automatically learns lists of positive sentiment phrases and negative situation phrases from sarcastic tweets. They show that identifying contrasting contexts yields improved recall for sarcasm recognition. More recently, Ghosh et al. [9] propose a reframing of sarcasm detection as a type of word sense disambiguation problem: given an utterance and a target word, identify whether the sense of the target word is literal or sarcastic.

Sarcasm as contextual phenomenon. Recently it has been observed that sarcasm requires some shared knowledge between the speaker and the audience; it is a profoundly contextual phenomenon [2]. Bamman et al. [2] use information about the authors, their relationship to the audience and the immediate communicative context to improve prediction accuracy. Rajadesingan et al. [27] adopt psychological and behavioral studies on when, why, and how sarcasm is expressed in communicative acts to develop a behavioral model and a set of computational features that merge user's current and past tweets as historical context. Joshi et al. [17] propose a framework based on the linguistic theory of context incongruity and introduce inter-sentential incongruity for sarcasm detection by considering the previous post in the discussion thread. Khattri et al. [19] present a quantitative evidence that historical tweets by an author can provide additional context for sarcasm detection. They exploit the author's past sentiment on the entities in a tweet to detect the sarcastic intent. Wang at al. [37] focus on message-level sarcasm detection on Twitter using a context-based model that leverages conversations, such as chains of tweets. They introduce a complex classification model that works over an entire tweet sequence and not on one tweet at a time. On the same direction, our work is based on the integration between linguistic and contextual features extracted from the analysis of visuals embedded in multimodal posts.

Sarcasm beyond text. Modern social media platforms allow to create multimodal forms of communication where audiovisual content integrates the textual utterance. Previous work [35] studied how different types of visuals are used in relation to irony in written discourse, and which pictorial elements contribute to the identification of verbal irony. Most scholars who looked at the relationship between verbal irony and images limited themselves to studying visual markers [1]. Usually a visual marker is either used to illustrate the literal meaning, or it may also exhibit incongruence with the literal evaluation of an ironic utterance (incongruence between the literal and intended evaluation). Following Kennedy [11], the image itself is usually considered not ironic; however, it may sometimes be important in deciding whether a verbal utterance is ironic or not. According to Verstraten [36], two types of elements play a role in the process of meaning-giving in the visual domain of static images. These include the *mise en scène* and *cinematographic* techniques. The mise en scène is concerned with the question of who and/or what is shown, cinematography deals with the question of how something is shown. Despite the similarities in the intent, our work shows few novel points: first of all, we analyze a large sample of non-

Platform	Text	Images
IG	Optional (up to 2,200 chars)	1
TU (photo)	Optional	1-10
TU (text)	Required	0 or more
TW	Required (up to 140 chars)	0 or more

Table 1: Text and image limitations.

Platform	#Posts	w/Text	w/Images	w/Both
IG	517,229	99.74%	100%	99.74%
TU	63,067	94.22%	45.99%	40.22%
TW	20,629	100%	7.56%	7.56%

Table 2: Presence of textual and visual components.

curated posts from three different social media platforms, while past work focuses mainly on curated content like advertisements, cartoons, or art. Moreover, to the best of our knowledge, we propose the first computational model that incorporates computer vision techniques to the automatic sarcasm detection pipeline.

Making sense of images. Recently, a number of research studies were devoted to combine visual and textual information, motivated by the progress of deep learning. Some approaches [21, 8] pursue a joint space for visual and semantic embedding, others consider how to generate captions to match the image content [24, 23], or how to capture the sentiment conveyed by an image [4, 38]. The most similar approach to our work is that of [31] which investigates the fusion of textual and image information to understand metaphors. A key aspect of our work is that it captures the relation between the visual and the textual dimensions as a whole, e.g., the utterance is not a mere description of an image, while in previous studies text is generally adopted to depict or model the content of an image.

3. DATA

To investigate the role images play in sarcasm detection, we collect data from three major social platforms that allow to post both text and images, namely Instagram (IG), Tumblr (TU) and Twitter (TW), using their available public APIs. Each of these platforms is originally meant for different purposes regarding the type of media to be shared. Whereas Instagram is an image-centric platform, Twitter is a microblogging network. Tumblr allows users to post different types of content, including "text" or "photo". Regardless of the post type, images (one or more) can be added to textual posts, and captions can be included in photo posts. The text and image restrictions and limitations for each platform are presented in Table 1.

The three platforms allow users to use hashtags to annotate the content, by embedding them in the text (Instagram, Twitter), or by adding them through a separate field (Tumblr). To collect positive (i.e., sarcastic) examples, we follow a hashtag-based approach by retrieving posts that include the tag *sarcasm* or *sarcastic*. This is a technique extensively used to collect sarcastic examples [9]. Additionally, and for all platforms, we filter out posts that are not in English, and remove retweets (Twitter) and reblogs (Tumblr) to keep the original content only and avoid duplicates.

Table 2 shows the distribution of posts with text, image(s), or both for each of the three platforms. Instagram is the platform where the the textual and visual modalities are most used in conjunction; in fact, almost the totality of posts have a caption accompanying the image. In contrast, less than 8% of the posts on Twitter contain images. Among the 63K Tumblr posts, 56.96% are of type "text", and 43.04% are of type "photo". This means that most of the photo posts contain also text (similar to Instagram, but without the limitation on the number of images), but very few of the text posts contain images (similar to Twitter, but without the character limitation).

Platform	#Words	#Emojis	#Tags
IG	10.77	0.37	7.44
TU	24.75	0.21	7.00
TW	9.45	0.29	1.96

Table 3: Average number of words, emojis and tags.

Filtering the data.

To clean up the data and build our final dataset we apply a series of four filters commonly used in literature [13, 6, 27]. First, we discard posts that do no contain any images, or whose images are no longer available by the time we collect the data; we then discard posts that contain mentions (@username) or external links (i.e., URLs that do not contain the platform name, or "t.co" or "twimg.com", in the case of Twitter), as additional information (e.g., conversational history, news story) could be required to understand the context of the message. We also discard posts where *sarcasm* or *sarcastic* is a regular word (not a hashtag), or a hashtag that is part of a sentence (i.e., if it is followed by any regular words), as we are not interested in messages that explicitly address sarcasm (e.g., "I speak fluent sarcasm."). Finally, we discard posts that might contain memes or *ecards* (e.g., tag set contains *someecards*), and posts whose text contains less than four regular words.

Final dataset. We randomly sample 10,000 posts from each platform to build our final dataset. Given the limitations of its public API, and the fact that less than 8% of the sarcastic posts have both text and images, only 2,005 were available for Twitter. We further clean up the data by removing internal links and the tags that we used to collect the samples (*sarcasm* and *sarcastic*). These posts are composed of two main aspects: a *textual* and a *visual* component. When we speak about the textual component, we are referring not only to the regular words that form the message, but also to *emojis* and *hashtags* that might be part of that message. These three elements (words, emojis and hashtags) are crucial for the interpretation of the post: while regular words are generally used to present the literal meaning, emojis and hashtags are commonly used to reveal the speaker's intended sentiment [16], or to share contextual cues with the audience to help decode the sarcasm.

Table 3 shows the average number of regular words, emojis and tags (after having removed *sarcasm*/*sarcastic*) per post. Due to its tight character limitation (which also accounts for the hashtags), Twitter is the platform with the shortest text and the lowest number of tags per post. While Tumblr posts are the longest, the average number of tags is similar to that of Instagram, which has in turn the highest tag-to-word ratio. Indeed, Instagram users seem to express

heavily through hashtags, especially compared to Twitter users, whose posts have a similar average word count. Both platforms also have a similar emoji-to-word ratio, which is much lower on Tumblr. The fact that there is a character limitation for both Instagram and Twitter might justify the usage of emojis, which are compact representations of concepts and reactions that would be much more verbose if expressed in words.

Finally, we collect 10,000 negative examples from each platform (2,005 from Twitter, to keep the dataset balanced) by randomly sampling posts that do **not** contain *sarcasm* or *sarcastic* in either the text or the tag set. These negative posts are subject to the same processing described above, when applicable. To verify that there are no relevant topical differences between the positive and the negative sets that could correlate with the presence/absence of sarcastic cues, we manually examined a sample of positive and negative posts from each platform. We did not observe such differences; however, we did find some recurring topics in the positive set, such as weather, food, fashion, etc., but these topics were also found in the negative set, only along with non-sarcastic observations (e.g., a picture of a greasy slice of pizza would be captioned as "healthy" in the positive set, but as "unhealthy" in the negative set). This might indicate that the range of topics in the positive set is more limited, but there is a clear overlap with those in the negative set.

4. CHARACTERIZING THE ROLE OF IMAGES IN SARCASTIC POSTS

As presented in Section 1, there are two main elements to a sarcastic utterance: the *context* and the *meaning* or *sentiment*. Detecting sarcasm—at a human level—involves evaluating to what extent the intended meaning corresponds to a declared or expected response. If this literal meaning does not agree with the one implied, the utterance will be perceived as sarcastic. In the following sections, we will analyze what role text (i.e., words, emojis and tags) and images play in the conception of sarcasm.

4.1 Defining a Categorization

To understand what role images play with respect to these two elements, three of the authors independently annotate a set of 100 randomly sampled positive posts from each platform. The question we are looking to answer is: *Is the image **necessary** to find the post sarcastic?* To answer that, we first identify the posts whose sarcastic nature can be positively determined by just looking at the text. This text, as explained in Section 3, can include words, emojis and tags. In many examples, emojis reveal the intended sentiment (in contrast to the literal sentiment presented in the regular text). Hashtags are generally useful to provide context, but can also be used to expose the sentiment. Regardless of whether the sarcastic tone is clear from the text or not, the image can still provide useful clues to understand the intended meaning. The posts where the intended meaning can **not** be inferred from the text alone are precisely what we are looking for. In these cases, the image turns out to be necessary to interpret the post, providing a depiction of the context, or visual clues to unravel the implied sentiment.

Table 4 summarizes the four possible roles of text and image. We will refer to the category that represents the combination of the two cases to the left as `Text Only`, as

Does the **IMAGE** help? \ Is the **TEXT** enough?	Yes	No
Yes	The text is clearly sarcastic; the image provides additional cues for better interpretability and engagement.	Both are needed to interpret the post. The clues to understand the intended meaning can be textual or visual.
No	The text is clearly sarcastic; the image does not provide any added value.	Post is **not** sarcastic.

Table 4: Roles of text and image in sarcastic posts.

the text from the posts belonging to it should be enough to understand the implied sarcasm. Figures 2(a) and 2(b) are instances of this category. The posts from the top-left case represent a subset of this category, where the image is somewhat redundant, but could replace or augment some of the textual clues. For instance, the image in Figure 2(b) would have been necessary if the tags *snow* and *winter* were not part of the text. In this case, also the emojis reveal the implied sentiment, which makes it unnecessary to infer that snow on a spring day is not "beautiful" or "nice", and that people are not supposed to wear shorts in such weather.

The top right case corresponds to the category that we will call `Text+Image`, where both modalities are required to understand the intended meaning. Figure 2(c) belongs to this category: the image depicts the context that the text refers to. Rather than a sentiment, the text presents an observation ("crowds of people") that is the opposite of what is shown in the picture (the room is empty). It is worth noting that, regardless of the category, many times the image itself contains text. In this case, the motivation to use an image instead of plain text is generally to provide additional information about the context of this text (e.g., a chat conversation, a screenshot, a street sign, and so on). Figure 2(a) is an example of this case.

4.2 Building a Ground Truth for Sarcasm

The data collection process described in Section 3 relies on the ability of the authors to self-annotate their posts as sarcastic using hashtags. Training a sarcasm detector on noisy data is a commonly used approach in literature, especially when that data comes from social media platforms. However, what the audience perceives as sarcastic is not always aligned with the actual intention of the speakers. Our goal is to create a curated dataset of multimodal posts whose sarcastic nature has been agreed on by both the author and the readers, and where both the textual and visual components are required to decode the sarcastic tone. To do that, we use CrowdFlower,[2] a large crowdsourcing platform that distributes small, discrete tasks to online contributors. The two goals of this annotation task are: 1) characterize the distribution of posts with respect to the categories defined in Section 4.1, and evaluate the impact of visuals as a source for context for humans; and 2) identify truly sarcastic posts by validating the authors' choice to tag them as such.

Task interface and setup. We focus only on the two main

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

(a) (b) (c)

Figure 2: Examples of sarcastic posts.

categories of interest, `Text Only` and `Text+Image`, and create two independent tasks. In the first task, only the text (including the tags and emojis) is shown to the annotator, along with the question "Is this text sarcastic?". The goal is to identify which posts belong to the `Text Only` category, i.e., posts where the textual component is enough to decode the sarcasm, and the image has a complementary role. We select 1,000 positive posts for this task, using the filters defined in Section 3. These posts are randomly sampled from the original sources, with no overlap with the dataset presented in that Section. We collect 5 annotations for each post, where the answer to the question can be "Yes" (text is sarcastic), "No" (text is **not** sarcastic) or "I don't know".

For the second experiment, we take only those posts that have been marked as non-sarcastic by the majority of the annotators on the first task (i.e., we discard the posts that belong to the `Text Only` category). Now we present both the textual and visual components, with the question "Is this post sarcastic?", and the same possible answers as before. Again, we collect 5 annotations per post.

The reason we run two independent experiments is to keep the tasks as simple as possible, and to guarantee that the judgment of the annotators is not affected by the knowledge that some information is missing. On the first task, annotators are not aware that the posts originally had one or more images, and are asked to judge them under that impression (same as a text-only based detector would do). If we did a two-step experiment instead, annotators would learn about the missing image(s) after having annotated the very first post, which would invite them to answer "I don't know" based on that indication. We run these experiments for both Instagram and Tumblr. Given the limited amount of data that we were able to collect for Twitter, and the fact that only a small percentage of the posts are actually multimodal, we do not build a gold set for this platform.

Quality control and inter-rater agreement. *Test Questions* (also called *Gold Standard* in CrowdFlower jargon) are curated job units that are used to test and track the contributor's performance and filter out bots or unreliable contributors. To access the task, workers are first asked to correctly annotate a set of *Test Questions* in an initial *Quiz Mode* screen, and their performance is tracked throughout the experiment with *Test Questions* randomly inserted in every task, disguised as normal units.

Judgments from contributors whose accuracy on the *Test Questions* is less than 78% are discarded and marked as not trusted.

Task	Matching%		Fleiss' κ	
	IG	TU	IG	TU
`Text Only (task 1)`	80.36	76.11	0.38	0.28
`Text+Image (task 2)`	74.65	86.40	0.21	0.23

Table 5: Inter-rater agreement.

To assess the quality of the collected data, we measure the level of agreement between annotators (see Table 5). *Matching%* is the percentage of matching judgments per object. For both experiments, the agreement is solid, with an average value around of 80%. However, the ratio of matching votes does not capture entirely the extent to which agreement emerges. We therefore compute the standard *Fleiss'* κ, a statistical measure for assessing the reliability of the agreement between a fixed number of raters. Consistently, the *Fleiss'* κ shows a Fair level [22] of agreement where, as expected, the second experiment reaches a lower agreement due to its intrinsic subjectivity and difficulty, even for human annotators [3].

Category	IG	TU
Not sarcastic	24.8%	31.9%
`Text Only`	37.8%	23.6%
`Text+Image`	**37.4%**	**44.5%**
D-80	19.1%	19.7%
D-100	8.6%	14.1%

Table 6: Percentage of posts in each category. The D-80 and D-100 subclasses refer to, respectively, posts where at least 80% or the totality of the annotators agree on the sarcastic nature of the post.

Results. Table 6 shows the distribution of the 1,000 posts with respect to the categories described in Section 4.1. For over 60% of the posts (62.20% for Instagram, 76.40% for Tumblr) the text alone (task 1) is **not** enough to determine

whether they are sarcastic or not. However, when those posts are shown with their visual component (task 2), more than half (60.13% for Instagram, 58.25% for Tumblr) are actually annotated as sarcastic, i.e., these posts were **misclassified** as non-sarcastic by the annotators on the first task, so the contribution of the image is crucial. It is interesting to note that a non-negligible fraction of the data (24.80% for Instagram, 31.90% for Tumblr) was not perceived as sarcastic by the majority of the annotators, which highlights the existing gap between the authors' interpretation of sarcasm and that of the readers, and the amount of noise we can expect in the dataset. In summary, the majority of the annotators found that both the text and the image are necessary to correctly evaluate the tone of the post in more than one third of the examples (37.40% for Instagram, 44.50% for Tumblr). Among these, 51.07% of the Instagram posts and 44.27% of the Tumblr posts were agreed to be sarcastic by at least 80% of the annotators (D-80), and 22.99% (IG) and 31.69% (TU) were unanimously declared sarcastic (D-100).

5. AUTOMATED METHODS FOR SARCASM DETECTION

We investigate two automatic methods for multimodal sarcasm detection. The first, a linear Support Vector Machine (SVM) approach, has been commonly used in prior work, though this prior work has relied on features extracted mainly from the text of the post (or set of posts). In our proposal, we combine a number of NLP features with visual features extracted from the image. The second approach relies on deep learning to fuse a deep network based representation of the image with unigrams as textual input. For both of these approaches, we evaluate the individual contributions of the respective textual and visual features, along with their fusion, in Section 6.

5.1 SVM Approach

For all experiments within this approach, we train a binary classification model using the sklearn toolkit[3] with its default settings.[4]

NLP Features. Our goal here is to replicate the prior art in developing a strong baseline composed of NLP features from which to investigate the impact that images have in detecting sarcasm. We adopt features commonly found in the literature: lexical features which measure aspects of word usage and frequency, features which measure the sentiment and subjectivity of the post, and word sequences (n-grams). We also make use of word embeddings, which has seen limited application to this task, save for a few works, such as [10], but has been used as a strong baseline in the sister task of sentiment analysis [7]. Finally, we select some of our best performing features and create a combination feature class. A description of each class is listed below:

- **lexical**: average word length, average word log-frequency according to the Google 1TB N-gram corpus,[5] number of contractions in sentence, average formality score as computed in [26].
- **subjectivity**: subjectivity and sentiment scores as computed by the TextBlob module,[6] number of passive constructions, number of hedge words, number of first person pronouns, number of third person pronouns.
- **n-grams**: unigrams and bigrams represented as one-hot features.
- **word2vec**: average of word vectors using pre-trained word2vec embeddings [25]. OOV words are skipped.
- **combination**: n-grams, word2vec and readability features (these include length of post in words and characters, as well as the Flesch-Kincaid Grade level score [20]).

Text is tokenized using nltk.[7] In addition, we treat hashtags in Instagram and Twitter, and tags in Tumblr, as well as emojis, as part of the text on which the features are derived from.

Visual Semantics Features (VSF). A key module to detect sarcasm is to understand the semantics in images. We employ the visual semantics models from Yahoo Flickr Creative Commons 100M (YFCC100M) [33], which include a diverse collection of complex real-world scenes, ranging from 200,000 street-life-blogged photos by photographer Andy Nystrom to snapshots of daily life, holidays, and events. Specifically, the semantics models were built with an off-the-shelf deep convolutional neural network using the Caffe framework [14], and the penultimate layer of the convolutional neural network output as the image-feature representation for training classifiers for 1,570 concepts which are popular in YFCC100M. Each concept classifier is a binary support vector machine, for which positive examples were manually labeled based on targeted search/group results, while the negatives drew negative examples from a general pool. The classifiers cover a diverse collection of visual semantics in social media, such as people, animals, objects, foods, architecture, and scenery, and will provide a good representation of image contents. Examples of concepts include terms such as "head", "nsfw", "outside", and "monochrome". In our experiments, we use the output of the content classifiers as one-hot features for the SVM regression model. Essentially, if a concept is detected, no matter what its associated confidence score, we treat it as a one-hot feature.

Multimodal Fusion. We concatenate the textual and visual features into a long vector, and once again use the linear SVM to train the fusion model. Previous research suggests that linear SVMs are fit for text classification [15], and our experiments find that linear SVM works very robustly to combine different kinds of features.

5.2 Deep Learning Approach

Adapted Visual Representation (AVR). The visual semantics classifiers described in the previous section are limited by a fixed vocabulary. To get a stronger visual representation, we follow the work in [28] and [18] that adopt a deep neural network. We borrow a model trained on ImageNet exactly from [5], which is based on roughly one million

[3]http://scikit-learn.org/

[4]We acknowledge that performance could be improved by experimenting with different parameters and kernels, however, our focus is not on optimizing for the best sarcasm detection system, but rather to construct a framework with which to show that visual features can complement textual features.

[5]https://catalog.ldc.upenn.edu/LDC2006T13

[6]https://textblob.readthedocs.org/en/dev/

[7]http://www.nltk.org/

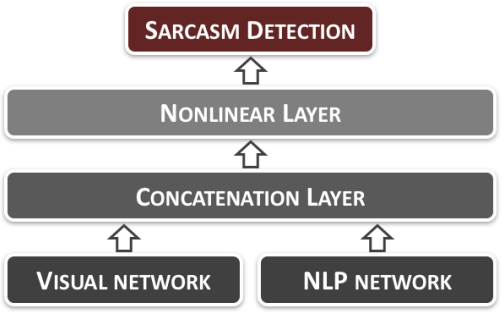

Figure 3: Network structure of our model. The visual network in the figure is initialized with the model weights in [5] trained on ImageNet.

images annotated with 1,000 object classes. There are originally seven layers in the model, but we remove the last layer of 1,000 neurons which correspond to the objects in ImageNet. The second to last layer has 4,096 neurons, which we will use to fine-tune with sarcastic and non-sarcastic data.

Textual Features. If we were to use all the NLP features in Section 5.1, our deep learning framework would quickly overfit given the limited size of the training set. As a consequence, a subset of the textual features were used in this fusion method. The NLP network is a two two layer perceptron based on unigrams only. The size of the first layer of the NLP network is the size of the unigram vocabulary for every platform. We employ a hidden layer in the NLP network with 512 hidden neurons, which is comparable with the number of neurons in the AVR.

Multimodal Fusion via Deep Network Adaptation. Figure 3 illustrates the neural network adaptation framework. We initialize a network with fixed image filters from the ImageNet model and random weights in other layers, and adapt it to our data. This adaption framework works with the deep CNN trained on ImageNet. The concatenation layer has 4,608 neurons. We use the rectify function as the activation function on all the nonlinear layers except for the last layer, which uses softmax over the two classes (sarcastic vs. non-sarcastic). Since in practice it is hard to find the global minimum in a deep neural network, we use Nesterov Stochastic Gradient Decent with a small random batch (size = 128). We finish training after 30 epochs.

6. EVALUATION

We evaluate our two methods under the same conditions, and with two different evaluation settings. For the first evaluation, models are developed on the data as described in Section 3, where we train on 50% of the data and evaluate on the remaining 50%. Please recall that the three data sets are evenly split between sarcastic and non-sarcastic posts, with the Instagram and Tumblr data sets containing a total of 20K posts each, and Twitter totaling 4,050 posts. We call this the **Silver Evaluation**, since the data is dependent on the authors correctly labeling their posts as sarcastic. As we saw in Table 6, 24.8% and 31.8% of Instagram and Tumblr posts marked by the authors as sarcastic are actually not sarcastic. For both the SVM and deep learning methods, we show results for *Text-Only*, *Image-Only* and the fusion of both modalities.

Next, we evaluate the respective Instagram and Tumblr models on the crowd-curated data sets in Section 4.2 (henceforth **Gold Evaluation**). Unlike the evaluation on the silver sets, the models are tested on re-judged data, and thus are of much higher quality, though there are fewer examples.

We use accuracy as our evaluation metric, and the baseline accuracy is 50% since all sets are evenly split.

6.1 Fusion with SVM

6.1.1 Evaluation on Silver Set

Feature Set	IG	TU	TW
lexical	56.7	54.3	57.8
subjectivity	61.7	59.9	58.3
1,2-grams	80.7	80.0	78.6
word2vec	74.9	73.6	75.3
combination	81.4	80.9	**80.5**
VSF only	68.8	65.7	61.7
n-gram + VSF	81.7	80.6	79.0
combination + VSF	**82.3**	**81.0**	80.0

Table 7: Silver Set evaluation using SVM fusion.

We first evaluate the contribution of the individual NLP features from Section 5.1 on the three data sets, as shown in the first main block in Table 7. The top individual feature is n-gram (1- and 2-grams), roughly performing at close to 80% accuracy across all data sets. In fact, even though we use three disparate data sets, the performance figures for each feature are consistently the same as the the ranking of the features. This may suggest that users do not alter the way they use sarcasm across platforms, though the best way of testing this hypothesis would be to investigate whether models trained on one platform, e.g., Twitter, can approximate the performance found on the other platforms, e.g., Instagram, when models are trained on native data. Finally, merging several of the feature classes into one (*combination*) yields the best performance, exceeding 80% for all data sets.

Using only the visual semantics features (VSF) yields an accuracy around 65% across the data sets. This is more than 15 points lower than the best NLP models; however, we were surprised that such a simple feature class actually outperformed the lexical and subjectivity features, both of which have been used in prior NLP work for the sarcasm detection task.

Finally, we combine the visual semantics features with the two best performing NLP features, i.e., n-grams and the combination feature class (last two rows of Table 7). For all the three data sets, the model with n-grams + VFS outperformed the model solely trained on n-grams by a small margin. However, it was not better than using the combination features. When combining the visual features with the combination features, we achieve the highest performance in Instagram (82.3%) and Tumblr (81.0%). In Twitter, the fusion produces the second highest performance (80.0%) to the 80.5% yielded by combination features only. These results show that including simple, noisy image-related features can improve sarcasm detection, albeit by a small margin.

6.1.2 Evaluation on Gold Set

Next, we investigate how well our models perform on the curated gold sets in Instagram and Tumblr. For the sake of simplicity, we focus our NLP evaluation on just the two top performing feature classes: n-grams and combination.

Feature Set	D-50 N=374	D-80 N=191	D-100 N=86
1,2-grams	81.7	81.9	80.2
combination	81.7	82.5	80.2
VSF only	75.7	72.8	68.0
1,2-grams + VSF	**86.6**	**87.7**	**83.7**
comb. + VSF	84.8	85.3	80.8

Table 8: SVM evaluation on Instagram Gold Sets.

Feature Set	D-50 N=445	D-80 N=197	D-100 N=141
1,2-grams	88.3	84.8	84.0
combination	**88.8**	86.0	84.4
VSF only	70.7	73.1	73.8
1,2-grams + VSF	88.5	**87.8**	**89.7**
comb. + VSF	88.0	87.1	**89.7**

Table 9: SVM evaluation on Tumblr Gold Sets.

Table 8 shows the results for the different modalities in Instagram. For the NLP features, the combination and n-gram are tied for the 50% and 100% agreement conditions (D-50 and D-100), while combination narrowly outperforms its counterpart in the 80% condition (D-80). As in the previous silver results, using the VSF only causes a loss in performance of nearly 15 points. The best results come from fusing n-grams with VSF, yielding a performance improvement of about 5% on all three agreement levels. Interestingly, while combination + VSF was generally the best feature in the silver evaluation, it is the second best here.

The Gold Tumblr results in Table 9 show a similar pattern with Table 8: the combination features outperform the n-gram features by a small margin across all three agreement levels, and only using VSF results in a performance loss of around 15 points accuracy compared to combination. We see the best performance when fusing the NLP and VSF features. At the 80% agreement level, n-gram + VSF yields a performance of 87.8%, which outperforms the best non-fusion performance by 1.8 points (86.0%). At the 100% agreement level, both fusion sets perform at 89.7%, a 5% point improvement. However, at the lower agreement rate (50%), the best performing fusion method just narrowly misses the combination method (88.5% to 88.8%).

The main message from both the silver and gold evaluations is that incorporating simple features which describe the image in a very traditional framework can improve performance. In general, the best performance comes not from fusing VSF with combination features, but rather with n-grams. We speculate that this may be due to the mismatch between the silver and gold sets. We do note that in some cases the performance improvement was small or non-existent. This is partially due to the noisiness of the data, the high baseline set by the NLP features, and also the accuracy of the VSF features, which can be viewed as hypotheses of what the classifier believes is present in the photo, even if weakly present.

6.2 Fusion with Deep Network Adaptation

Next, we evaluate our deep learning approach on our silver and gold sets. We additionally evaluate the model with image (AVR) and text (unigram) features only, for which the concatenation layer (see Figure 3) still exists but has no effect with single modality input. The three models use the same learning rates.

6.2.1 Evaluation on Silver Set

Feature Set	IG	TU	TW
1-grams	71.0	65.3	54.1
AVR only	73.8	69.2	68.7
1-grams + AVR	**74.2**	**70.9**	**69.7**

Table 10: Silver Set evaluation using DNA fusion.

Table 10 shows the the evaluation on the silver set. It is easy to see that fusing the textual and image signals together provides the best performance across all three sets, ranging from 74.2% in Instagram to 69.7% in Twitter. That confirms our hypothesis that the visual aspect plays a role in the detection of sarcasm.

Another interesting phenomenon is that the image-only network outperforms the visual semantics features consistently in all three platforms: 73.8% vs. 68.8% in Instagram, 69.2% vs. 65.7% in Tumblr, and 68.7% vs. 61.7% in Twitter. This suggests that the adapted deep CNN better captures the diversity of sarcastic images. On the other hand, our text-based network is worse than the text models using SVM. The reason is mainly because our text network does not use bigrams or higher dimensional features. Since the visual semantics features are not fine-tuned, the simpler fusion by SVM method does not overfit the training set. As a result, all state-of-the-art NLP features described in Section 5.1 can be used in this model.

Among the three platforms, the performance in Twitter is lower than in the other two. We believe that this is mainly due to the small amount of training data (2,000 posts), which is an issue for deep learning. Also, given that Twitter is mostly a textual platform (especially compared to the more image-centric Instagram and Tumblr), the weaker textual baseline seems to fail to capture the nuances of sarcasm used in this platform.

6.2.2 Evaluation on Gold Set

Feature Set	D-50 N=374	D-80 N=191	D-100 N=86
1-grams	69.7	67.7	63.1
AVR only	77.0	74.6	74.8
1-grams + AVR	**77.8**	**78.4**	**77.6**

Table 11: DNA evaluation on Instagram Gold Sets.

Feature Set	D-50 N=445	D-80 N=197	D-100 N=141
1-grams	68.4	65.8	64.6
AVR only	75.8	74.6	**75.5**
1-grams + AVR	**77.6**	**75.6**	74.7

Table 12: DNA evaluation on Tumblr Gold Sets.

Our gold results show a similar pattern. In the Tumblr set, the fusion of text and image yields the best performance over D-50 and D-80, but is narrowly behind just using the image on D-100. In the Instagram set, the fusion of text and images yields the best performance in all three platforms. Since the text feature is limited, the performance of deep network adaptation is not as competitive as the SVM based fusion method. However, we think the performance of deep neural network adaption will be improved with more training examples.

7. CONCLUSIONS

To the best of our knowledge, this work represents the first empirical investigation on the impact of images for sarcasm detection in social media. In particular, we first investigate the role of images, and quantitatively show that humans use visuals as situational context to decode the sarcastic tone of a post. The collected and annotated data will be shared with the research community. Second, we show that automatic methods for sarcasm detection can be improved by taking visual information into account. Finally, while most previous work has focused on the study of textual utterances on Twitter, our research shows breadth by tackling two other popular social media platforms: Instagram and Tumblr.

We propose two types of multimodal fusion frameworks to integrate the visual and textual components, and we evaluate them across three social media platforms with heterogeneous characteristics. With the use of visual semantics features, we observe an improved performance for the noisy dataset in the case of Instagram (the most image-centric platform), while the impact of images in Tumblr and Twitter was not perceived as relevant. We argue that this behavior is due to their text-centric nature. In the case of curated data though, we observe higher predictive accuracy across all the platforms, and across almost all of the agreement levels, which suggests that the visual component plays an important role when human judgments are involved.

By using deep network adaptation, we show a consistent increment in performance across the three platforms. Also in this case, Instagram was the platform that reached the highest accuracy. We have pointed out the weak performance of the textual features used in the deep learning approach. The challenges that prevent us from using more advanced textual features (such as those used in the SVM model) are two-fold: 1) given the limited size of the training set, the network adaptation method suffers from overfitting; adding new features does not help when the fusion network can get almost perfect accuracy on the training set; and 2) a higher dimensionality brings difficulties for a fast neural network training due to the limitations of the GPU memory. Collecting more training data should, at the very least, address the overfitting issue.

Images can be thought of as another form of contextual clue, much like the role of previous tweets by a user or the overall sarcasm levels of a discussion thus far. In our future work, we wish to build a model which integrates all these contextual clues within our framework to assess which ones have the largest impact per platform. We are also interested in including visual sentiment frameworks in the evaluation of the sarcastic tone.

8. ACKNOWLEDGMENTS

This work is partially supported by the project "ExceptionOWL: Nonmonotonic Extensions of Description Logics and OWL for defeasible inheritance with exceptions", Progetti di Ateneo Università degli Studi di Torino and Compagnia di San Paolo, call 2014, line "Excellent (young) PI".

9. REFERENCES

[1] S. Attardo, J. Eisterhold, J. Hay, and I. Poggi. Multimodal markers of irony and sarcasm. *Humor-international Journal of Humor Research*, 16:243–260, 2003.

[2] D. Bamman and N. A. Smith. Contextualized sarcasm detection on twitter. In M. Cha, C. Mascolo, and C. Sandvig, editors, *Proc. of the Ninth Int. Conference on Web and Social Media, ICWSM*, pages 574–577. AAAI Press, 2015.

[3] F. Barbieri, H. Saggion, and F. Ronzano. Modelling sarcasm in twitter, a novel approach. In *Proc. of the Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 50–58, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

[4] D. Borth, T. Chen, R. Ji, and S.-F. Chang. Sentibank: Large-scale ontology and classifiers for detecting sentiment and emotions in visual content. In *Proc. of the ACM Int. Conference on Multimedia*, MM '13, pages 459–460, New York, NY, USA, 2013. ACM.

[5] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: delving deep into convolutional nets. In *BMVC*, 2014.

[6] D. Davidov, O. Tsur, and A. Rappoport. Semi-supervised recognition of sarcastic sentences in twitter and amazon. In *Proc. of the Conference on Computational Natural Language Learning*, CoNLL '10, pages 107–116, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[7] M. Faruqui and C. Dyer. Non-distributional word vector representations. In *Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)*, volume 2, pages 464–469, Beijing, China, July 2015. Association for Computational Linguistics.

[8] A. Frome, G. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov. Devise: A deep visual-semantic embedding model. In *Advances In Neural Information Processing Systems, NIPS*, 2013.

[9] D. Ghosh, W. Guo, and S. Muresan. Sarcastic or not: Word embeddings to predict the literal or sarcastic meaning of words. In L. Màrquez, C. Callison-Burch, J. Su, D. Pighin, and Y. Marton, editors, *EMNLP*, pages 1003–1012. The Association for Computational Linguistics, 2015.

[10] D. Ghosh, W. Guo, and S. Muresan. Sarcastic or not: Word embeddings to predict the literal or sarcastic meaning of words. In *Proc. of the Conference on Empirical Methods in Natural Language Processing*, pages 1003–1012, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[11] R. Gibbs. *The Cambridge Handbook of Metaphor and Thought*. Cambridge Handbooks in Psychology. Cambridge University Press, 2008.

[12] R. Gibbs and H. Colston. *Irony in Language and Thought: A Cognitive Science Reader*. Lawrence Erlbaum Associates, 2007.

[13] R. González-Ibáñez, S. Muresan, and N. Wacholder. Identifying sarcasm in twitter: A closer look. In *Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)*, volume 2 of *HLT '11*, pages 581–586, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[14] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[15] T. Joachims. Text categorization with suport vector machines: Learning with many relevant features. In *Proceedings of the 10th European Conference on Machine Learning*, ECML '98, pages 137–142, London, UK, UK, 1998. Springer-Verlag.

[16] A. Joshi, P. Bhattacharyya, and M. J. Carman. Automatic sarcasm detection: A survey. *CoRR*, abs/1602.03426, 2016.

[17] A. Joshi, V. Sharma, and P. Bhattacharyya. Harnessing context incongruity for sarcasm detection. In *Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)*, volume 2, pages 757–762. The Association for Computer Linguistics, 2015.

[18] S. Karayev, A. Hertzmann, H. Winnemoeller, A. Agarwala, and T. Darrell. Recognizing image style. *BMVC*, 2014.

[19] A. Khattri, A. Joshi, P. Bhattacharyya, and M. Carman. Your sentiment precedes you: Using an author's historical tweets to predict sarcasm. In *Proc. of the Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 25–30, Lisboa, Portugal, September 2015. Association for Computational Linguistics.

[20] J. P. Kincaid, R. P. Fishburne Jr., R. L. Rogers, and B. S. Chissom. Derivation of new readability formulas (automated readability index, fog count and Flesch reading ease formula) for navy enlisted personnel. Technical report, DTIC Document, 1975.

[21] R. Kiros, R. Salakhutdinov, and R. S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *CoRR*, abs/1411.2539, 2014.

[22] J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 1977.

[23] L. Ma, Z. Lu, L. Shang, and H. Li. Multimodal convolutional neural networks for matching image and sentence. *CoRR*, abs/1504.06063, 2015.

[24] J. Mao, W. Xu, Y. Yang, J. Wang, and A. L. Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). *CoRR*, abs/1412.6632, 2014.

[25] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[26] E. Pavlick and A. Nenkova. Inducing lexical style properties for paraphrase and genre differentiation. In *Proc. of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 218–224, Denver, Colorado, May–June 2015. Association for Computational Linguistics.

[27] A. Rajadesingan, R. Zafarani, and H. Liu. Sarcasm detection on twitter: A behavioral modeling approach. In *Proc. of the ACM Int. Conference on Web Search and Data Mining*, WSDM '15, pages 97–106, New York, NY, USA, 2015. ACM.

[28] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. CNN features off-the-shelf: an astounding baseline for recognition. *CVPR DeepVision workshop*, 2014.

[29] A. Reyes, P. Rosso, and T. Veale. A multidimensional approach for detecting irony in twitter. *Lang. Resour. Eval.*, 47(1):239–268, Mar. 2013.

[30] E. Riloff, A. Qadir, P. Surve, L. D. Silva, N. Gilbert, and R. Huang. Sarcasm as contrast between a positive sentiment and negative situation. In *EMNLP*, pages 704–714. ACL, 2013.

[31] E. Shutova, D. Kiela, and J. Maillard. Black holes and white rabbits: Metaphor identification with visual features. In *Proc. of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 160–170, San Diego, California, June 2016. Association for Computational Linguistics.

[32] J. Tepperman, D. Traum, and S. S. Narayanan. "yeah right": Sarcasm recognition for spoken dialogue systems. In *Proc. of InterSpeech*, pages 1838–1841, Pittsburgh, PA, Sept. 2006.

[33] B. Thomee, B. Elizalde, D. A. Shamma, K. Ni, G. Friedland, D. Poland, D. Borth, and L.-J. Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016.

[34] O. Tsur, D. Davidov, and A. Rappoport. Semi-Supervised Recognition of Sarcastic Sentences in Online Product Reviews. In M. Hearst, W. Cohen, and S. Gosling, editors, *Proc. of the Int. Conference on Weblogs and Social Media (ICWSM-2010)*. The AAAI Press, Menlo Park, California, 2010.

[35] T. Veale, K. Feyaerts, and C. Forceville. *Creativity and the Agile Mind: A Multi-Disciplinary Study of a Multi-Faceted Phenomenon*. Applications of Cognitive Linguistics [ACL]. De Gruyter, 2013.

[36] P. Verstraten. *Film Narratology : film narratives his primary focus, while noting*. University of Toronto Press, 2006.

[37] Z. Wang, Z. Wu, R. Wang, and Y. Ren. Twitter sarcasm detection exploiting a context-based model. In J. Wang, W. Cellary, D. Wang, H. Wang, S. Chen, T. Li, and Y. Zhang, editors, *Web Information Systems Engineering - WISE 2015 - 16th Int. Conference, Miami, FL, USA, November 1-3, 2015, Proc., Part I*, volume 9418 of *Lecture Notes in Computer Science*, pages 77–91. Springer, 2015.

[38] Q. You, J. Luo, H. Jin, and J. Yang. Robust Image Sentiment Analysis using Progressively Trained and Domain Transferred Deep Networks. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2015.