Digital Leafleting: Extracting Structured Data from Multimedia Online Flyers

Emilia Apostolova
BrokerSavant Inc.
2506 N. Clark St, 415
Chicago, IL 60614, USA
emilia@brokersavant.com

Payam Pourashraf
DePaul University
243 S Wabash Ave
Chicago, IL 60604, USA
ppourash@cdm.depaul.edu

Jeffrey Sack
BrokerSavant Inc.
2506 N. Clark St, 415
Chicago, IL 60614, USA
jeff@brokersavant.com

Abstract

Marketing materials such as flyers and other infographics are a vast online resource. In a number of industries, such as the commercial real estate industry, they are in fact the only authoritative source of information. Companies attempting to organize commercial real estate inventories spend a significant amount of resources on manual data entry of this information. In this work, we propose a method for extracting structured data from free-form commercial real estate flyers in PDF and HTML formats. We modeled the problem as text categorization and Named Entity Recognition (NER) tasks and applied a supervised machine learning approach (Support Vector Machines). Our dataset consists of more than 2,200 commercial real estate flyers and associated manually entered structured data, which was used to automatically create training datasets. Traditionally, text categorization and NER approaches are based on textual information only. However, information in visually rich formats such as PDF and HTML is often conveyed by a combination of textual and visual features. Large fonts, visually salient colors, and positioning often indicate the most relevant pieces of information. We applied novel features based on visual characteristics in addition to traditional text features and show that performance improved significantly for both the text categorization and NER tasks.

1 Introduction

Digital flyers are the preferred and sometimes only method of conveying offerings information in a number of broker-based industries. Such industries typically lack a centralized database or an established source of information. Organizing such content typically involves manual data entry, an expensive and labour intensive effort. Further challenge is that available offerings constantly change and manually entered data often results in out-dated inventories.

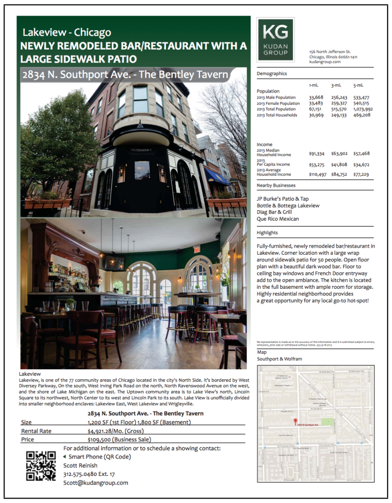
In particular, the commercial real estate industry in the US (unlike residential real estate[1]) does not have a centralized database or an established source of information. A number of commercial real estate inventories collect commercial real estate data using information from flyers, contacting brokers, or visiting physical sites[2]. The data is manually entered in structured form. At the same time, inventories change on a daily basis. As a result, the collected information is typically sparse and often outdated. In fact, commercial real estate brokers often need to rely on networking and chance in preference to consulting third party listing databases.

While brokers do not often update third party inventory databases, they do create marketing materials (usually PDF flyers and/or HTML emails/web pages) that contain all relevant listing information. Virtually all commercial real estate offerings come with a publicly available marketing material that contains all relevant listing information. Figures 1 and 2 show two typical commercial real estate flyers.

[1] The US Multiple Listing Services (MLS), governed by the National Association of Realtors, represents the US residential real estate.

[2] LoopNet, subsidiary of CoStar Group Inc., is the most heavily trafficked online commercial real estate inventory.

Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL, pages 283–292,
Denver, Colorado, May 31 – June 5, 2015. ©2015 Association for Computational Linguistics



Lakeview - Chicago

NEWLY REMODELED BAR/RESTAURANT WITH A LARGE SIDEWALK PATIO

2834 N. Southport Ave. - The Bentley Tavern

KG KUDAN GROUP

156 North Jefferson St.
Chicago, Illinois 60661-1411
kudangroup.com

Demographics

	1-mi.	3-mi.	5-mi.
Population			
2013 Male Population	33,668	256,243	533,477
2013 Female Population	33,483	259,327	540,515
2013 Total Population	67,151	515,570	1,073,992
2013 Total Households	30,969	249,133	469,208
Income			
2013 Median Household Income	$91,334	$63,902	$57,468
2013 Per Capita Income	$53,275	$41,808	$34,672
2013 Average Household Income	$110,497	$84,752	$77,229

Nearby Businesses

JP Burke's Patio & Tap
Bottle & Bottega Lakeview
Diag Bar & Grill
Que Rico Mexican

Highlights

Fully-furnished, newly remodeled bar/restaurant in Lakeview. Corner location with a large wrap around sidewalk patio for 50 people. Open floor plan with a beautiful dark wood bar. Floor to ceiling bay windows and French Door entryway add to the open ambiance. The kitchen is located in the full basement with ample room for storage. Highly residential neighborhood provides a great opportunity for any local go-to hot-spot!

No representation is made as to the accuracy of this information and it is submitted subject to errors, omissions, prior sale or withdrawal without notice. 355 9-18-2013

Map
Southport & Wolfram

Lakeview

Lakeview, is one of the 77 community areas of Chicago located in the city's North Side. It's bordered by West Diversey Parkway, On the south, West Irving Park Road on the north, North Ravenswood Avenue on the west, and the shore of Lake Michigan on the east. The Uptown community area is to Lake View's north, Lincoln Square to its northwest, North Center to its west and Lincoln Park to its south. Lake View is unofficially divided into smaller neighborhood enclaves: Lakeview East, West Lakeview and Wrigleyville.

	2834 N. Southport Ave. - The Bentley Tavern
Size	1,200 SF (1st Floor) 1,800 SF (Basement)
Rental Rate	$4,921.28/Mo. (Gross)
Price	$109,500 (Business Sale)

For additional information or to schedule a showing contact:
◂ Smart Phone (QR Code)
Scott Reinish
312.575.0480 Ext. 17
Scott@kudangroup.com

Figure 1: An example of a commercial real estate flyer © Kudan Group Real Estate.

Our goal is to utilize this publicly available information and extract structured data that can be continuously updated for a reliable centralized database of offerings.

Commercial listing information is typically summarized in a structured form suitable for targeted property searches. The most important information consists of the various categories of the offering. For example, the transaction type (*sale*, *lease*, and/or *investment*), the property type (*industrial, retail, office*, etc.), the location of the property (its full geocoded address), the size of the property, the contact information of the brokers representing the property, etc.

This information is typically present in text form within the flyer. However, flyers and similar marketing materials are essentially multi-media documents. In addition to text, information is also conveyed by attributes such as font size, color, positioning, and images. For example, the listing address of the flyer on Figure 1 can be easily identified by its prominent color, size, and positioning (*2834 N. Southport Ave*, upper left corner). While the address of the broker firm shown in the same flyer is considered non-essential information and lacks such visual prominence (*156 North Jefferson St.*, upper right corner). In fact, it is very difficult and sometimes impossible to distinguish between the two address types when considering a text-only version of the flyer. Similarly, the transaction type (*For Sale*) of the property on Figure 2 is prominently shown in a large font and distinctive color. To account for the multi-media nature of the dataset, we attempt to combine textual and visual features for the task of automatic extraction of structured information from free-form commercial real estate flyers.

The problem of extracting structured data from flyers was modeled as text categorization and Named Entity Recognition (NER) tasks as described in Section *Problem Definition* below. Typically, both text categorization and NER approaches are applied to genres with exclusively text-based content (newswires, scientific publications, blogs and other social media texts). As a result, the feature space of NER and text categorization involves purely textual features: word attributes and characteristics, their contexts and frequencies. However, textual information in visually rich formats, such as PDF and HTML, is interlaced with typographic and other visually salient characteristics. In this study, we propose several novel features that take visual characteristics into account and show that performance improves significantly on both the text categorization and NER tasks.

2 Problem Definition

Given a commercial real estate flyer, our task is to extract structured information that can be used as input to a commercial real estate listing service. This information includes categories associated with the property (property type and transaction type) and a list of property attributes (address, space information, and broker information).

The task of identifying a list of categories was modeled as a text categorization task. The categories and associated types are summarized in Table 1. Both text categorization tasks (identifying the Transaction and Property Types) are multi-label classification tasks, i.e. multiple category labels can be assigned to each listing. For example, properties are often offered for both *sale* and *lease* and belong to both transaction types. Similarly, a retail building could offer an associated office unit and belongs to property types *retail* and *office*.

The task of identifying values of specific listing attributes was modeled as a Named Entity Recognition (NER) task. The various NER types and de-

Transaction Type	A listing can have one or more of the following transaction types: *sale, lease, investment*.
Property Type	A listing can have one or more of the following property types: *retail, office, industrial, land, multi-family*.

Table 1: Types and descriptions of flyer categories.

scriptions are summarized in Table 2. The named entities represent a typical set of attributes collected by commercial real estate listing services. They are 1) one or more brokers representing the property and their contact information; 2) the full address of the property broken down into individual address fields; 3) one or more spaces including their sizes and types (e.g. sizes of available units in a shopping mall, the sizes of a warehouse building and associated office building, etc.).

Broker Name Broker Email Broker Phone	The contact information of all listing brokers, including full name, email address, phone number.
Company Name	The brokerage company name.
Street City Neighborhood State Zip	The address information of the listing address including street or intersection, city, state, and, zip code.
Space Size and Type	Size and attributes of relevant spaces (e.g. *27,042 SF building, 4.44 acres site*, etc.); Includes the numeric value, unit of measure, whether the value is a part of a range (min or max) or exact value, as well as the space type *(unit, building,lot*); Excludes size information of non-essential listing attributes (e.g. basement size or parking lot size).

Table 2: Types and descriptions of named entities relevant to extracting listing information from commercial real estate flyers.

The problem of automatically extracting structured information from real estate flyers was then implemented as a combination of the text categorization and NER tasks.

3 Method

3.1 Dataset

The dataset consists of 2,269 commercial real estate flyers submitted to a listing service[3] over a period of one year. It represents over 450 US locations, over 90 commercial real estate companies and over 800 commercial real estate brokers. The flyers were generated using different tools and formats and represent a wide variety of styles. Typically, the same broker can represent properties of various categories and transaction types. The text categorization was evaluated using the full dataset of 2,269 flyers with 5-fold cross validation. For the NER task, we used 60 percent of the flyers for training (a total of 1,361 flyers), and the remaining 40 percent for testing (a total of 908 flyers).

All flyers (PDF and HTML) were converted to a common format (HTML)[4]. The flyers were converted to text using an HTML parser and extracting DOM[5] text elements while preserving some of the white space formatting. In some cases, text was presented inside embedded images within the flyer. Since the majority of flyers, however, were mostly in text format, OCR[6] was not used and text within images was discarded. The median number of characters, tokens, and sentences for flyers are 2106, 424, and 72 respectively.

3.2 Training Data Transformation

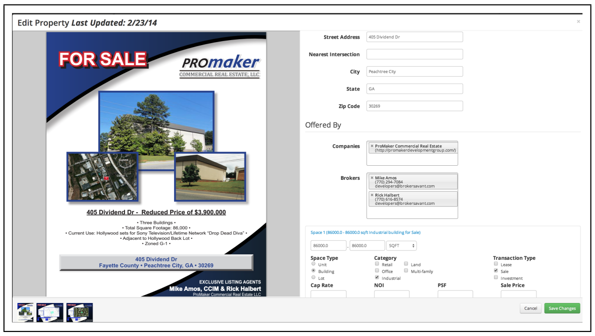
In previous work, we have created a tool used to annotate HTML flyers and evaluated the NER task on a subset of 800 manually annotated flyers. However, the manual annotation proved a laborious task (the same listing attribute typically appears multiple times in the document) and resulted in moderate inter-annotator agreement. In this work, instead of manually annotating the full set of 2,269 flyers, we used listing data entered by professional data en-

[3] © BrokerSavant Inc.

[4] PDFs were converted to HTML using the PDFTOHTML conversion program `http://pdftohtml.sourceforge.net/`. While the open-source tool occasionally misaligned text, performance was satisfactory and the use of more accurate commercial PDF-to-HTML5 conversion tool was deemed unnecessary.

[5] Document Object Model.

[6] Optical Character Recognition.



Edit Property *Last Updated: 2/23/14*

FOR SALE

PROmaker
COMMERCIAL REAL ESTATE, LLC

405 Dividend Dr - Reduced Price of $3,900,000

• Three Buildings •
• Total Square Footage: 86,000 •
• Current Use: Hollywood sets for Sony Television/Lifetime Network "Drop Dead Diva" •
• Adjacent to Hollywood Back Lot •
• Zoned G-1 •

405 Dividend Dr
Fayette County • Peachtree City, GA • 30269

EXCLUSIVE LISTING AGENTS
Mike Amos, CCIM & Rick Halbert
ProMaker Commercial Real Estate LLC

Street Address: 405 Dividend Dr

Nearest Intersection:

City: Peachtree City

State: GA

Zip Code: 30269

Offered By

Companies: × ProMaker Commercial Real Estate (http://promakerdevelopmentgroup.com/)

Brokers: × Mike Amos (770) 294-7084 developers@brokersavant.com
× Rick Halbert (770) 616-8574 developers@brokersavant.com

Space 1 (86000.0 - 86000.0 sqft Industrial building for Sale)

86000.0 - 86000.0 SQFT

Space Type: ○ Unit ● Building ○ Lot

Category: □ Retail □ Office ☑ Industrial □ Land □ Multi-family

Transaction Type: □ Lease ☑ Sale □ Investment

Cap Rate NOI PSF Sale Price

Figure 2: An example of a commercial real estate flyer and manually entered listing information © ProMaker Commercial Real Estate LLC, © BrokerSavant Inc.

try staff employed by the listing service[7]. Figure 2 shows an example of a real estate flyer and the corresponding manually entered listing data.

To generate a dataset for the text categorization tasks we assigned the list of manually entered labels for transaction and property types to each flyer. For example, the flyer from Figure 2 was assigned to transaction type *sale* and property type *industrial*.

To generate annotated data for the NER task, we had to convert the stand-alone listing information to annotated text in which each occurrence of the field values was marked with the corresponding entity type via string matching. The manually entered listing data, however, introduced some text variations and did not always match the text in the corresponding flyer. For example, the same street and intersection address could be expressed in a variety of ways (e.g. *'Westpark Drive and Main Street' vs 'Westpark Dr & Main St'; '123 North Main Road' vs '123 N Main'*, etc.). Similarly, broker names, phones, and company names could have a variety of alternative representations (e.g. *'Michael R. Smith' vs 'Mike Smith CCIM'; 'Lee and Associates' vs 'Lee & Associates of IL LLC'; '773-777-0000 ext 102' vs '773.777.0000x102'*, etc.). Lastly, space size information was always entered in square feet, while at the same time it could be expressed as both square feet and acres (with various precision points) in the corresponding flyer (e.g. *53796 sf, 1.235 acres, 1.23 acres*, etc.).

To account for the various ways in which an attribute value can be expressed in the text we hand-crafted a small set of rules and regular expressions that allowed us to find most of its alternative representations. In some cases, however, the listing value was not found in the corresponding flyer text. In the case of such a discrepancy, the flyer was simply discarded from the training set used for the corresponding named entity type. Such discrepancies could occur for several reasons. In some cases, the manually hand-crafted rules and regular expressions did not cover all possible variants in which the value could be expressed. On occasion, the text containing the attribute value was in image format (inside embedded images). We also noted a few instances of incorrectly entered manual data. As a result, only a portion of the training data (a total of 1,361 flyers) was used for the training of individual named entity types. We were able to automatically annotate 878 flyers used for training the address named entity recognizer (street or intersection, city, state, zip), 1145 flyers used for training the broker information named entity recognizer (broker name, phone, email, company) and 1242 flyers for training the space named entity recognizer (size and space type).

3.3 Data Pre-processing

As mentioned earlier, all flyers (PDF and HTML) were converted to a common format (HTML). An HTML parser was then used to create a text representation of the flyer. The text was tokenized and tokens were normalized (all tokens were converted to lower case and digits were converted to a common format).

As noted previously, data entry staff were able to quickly spot listing attributes of interest solely because of their visual characteristics. To account for such visual characteristics we included typographic and other visual features associated with tokens or text chunks for both the text categorization and NER tasks. Typographic and visual features were based on the computed HTML style attributes for each DOM element containing text.

Computing the HTML style attributes is a complex task since they are typically defined by a combination of CSS[8] files, in-lined HTML style attributes, and browser defaults. The complexities of style definition, inheritance, and overwriting are handled by

[7] © BrokerSavant Inc.

[8] Cascading Style Sheets.

browsers[9]. We used the Chrome browser to dynamically compute the style of each DOM element and output it as inline style attributes. To achieve this we programmatically inserted a javascript snippet that inlines the computed style and saves the new version of the HTML on the local file system utilizing the HTML5 *saveAs* interface[10]. We then normalized the style attribute values for font size, RGB color, and Y coordinate as described in the following sections.

3.4 Text Categorization

The text categorization task involves labeling all flyers with appropriate transaction types and property types as shown in Table 1. This is a multi-label classification task as in all cases a flyer can have more than one label (e.g. Transaction Type: *sale, lease*; Property Type: *retail, office*).

We applied a supervised Machine Learning approach to the task utilizing Support Vector Machines (SVM) (Vapnik, 2000) using the LibSVM library (Chang and Lin, 2011). SVM was a sensible choice as it has been shown to be one of the top performers on a number of text categorization tasks (Joachims, 1998; Yang and Liu, 1999; Sebastiani, 2002).

Category information such as the transaction type and property type are one of the key pieces of information in a flyer. However, they are not always explicitly mentioned in the flyer and in some cases the data entry person needs to read the full content of the flyer to infer the property type. For example, an *industrial* property might be inferred by a mention of a particular zoning category and description of loading docks; a *retail* property type might be inferred by mentions of retail neighbors (e.g. *Staples, Bed Bath and Beyond*, etc) and traffic counts; an *investment* property type can be inferred by description of NOI (net operating income) and Cap Rates (the ratio between the NOI and capital cost of a property), etc. At the same time, when present, terms indicating the transaction and property types typically appear prominently in large fonts. For example, the property type of the flyer shown on Figure 1 is prominently shown in large font (*Restaurant* indicates *retail* property type). Similarly, the transaction type of the flyer shown on Figure 2 is again prominently displayed in a large font (*For Sale*). The classifiers could then benefit from both the full text of the flyers, combined with some information of the visual prominence of individual words.

We used 'bag-of-words' representation (token unigrams) and modeled the task as a binary classification for each category label. As a term weighting scheme, we first used TF-IDF as one of the most common weighting schemes used for term categorization (Lan et al., 2005; Soucy and Mineau, 2005). This served as a performance baseline. To account for visually prominent characteristics of important document terms we also introduced a term weight that takes into account the relative font size of the term. As a measure of the relative font size, we used the percentile rank of the term font size, compared to all term font sizes in the document. For example, a weight of 0.9 is assigned to terms whose font size is greater than 90% of all tokens within the current document. The font size percentile was then used as a term weighting scheme (instead of TF-IDF). Table 3 summarizes the results of 5-fold cross validation using the full dataset of 2,269 flyers. We used a linear kernel model with the default parameters.

		TF-IDF	Font Size Pctl
Property type	P	79.57	85.04
	R	85.27	84.16
	F	82.32	**84.6**
Transaction type	P	87.56	89.64
	R	92.87	94.60
	F	90.14	**92.05**

Table 3: Results from applying SVM on the task of identifying flyer Property Types (*retail, office, industrial, land, multi-family*) and Transaction Types (*sale, lease, investment*). We used 'bag-of-words' representation (unigrams) applying two different term weight schemes: TF-IDF and the relative percentile rank of the term font size. P=precision, R=recall, F=f1 score.

In both text categorization tasks the Font Size Percentile term weight significantly outperformed the TF-IDF term weight scheme[11].

[9] We attempted to use an HTML renderer from the Cobra java toolkit `http://lobobrowser.org/cobra.jsp` to compute HTML style attributes. However, this renderer produced poor results on our dataset and failed to accurately compute the pixel location of text elements.

[10] `https://github.com/eligrey/FileSaver.js`

[11] The difference is statistically significant with p value $<$ 0.05% using Z-test on two proportions.

3.5 Named Entity Recognition

A supervised machine learning approach was then applied to the task of identifying the named entities shown in Table 2. The task was modeled as a **BIO** classification task, classifiers identify the **B**eginning, the **I**nside, and **O**utside of the text segments. We first used a traditional set of text-based features for the classification task. Table 4 lists the various text-based features used. In all cases, a sliding window including the 6 preceding and 6 following tokens was used as features.

Feature Name	Description
Token	A normalized string representation of the token. All tokens were converted to lower case and all digits were converted to a common format.
Token Orth	The token orthography. Possible values are lowercase (all token characters are lower case), all capitals (all token characters are upper case), upper initial (the first token character is upper case, the rest are lower case), mixed (any mixture of upper and lower case letters not included in the previous categories).
Token Kind	Possible values are word, number, symbol, punctuation.
Regex type	Regex-based rules were used to mark chunks as one of 3 regex types: email, phone number, zip code.
Gazetteer	Text chunks were marked as possible US cities or states based on US Census Bureau city and state data. www.census.gov/geo/maps-data/data/gazetteer2013.html.

Table 4: List of text-based features used for the NER task. A sliding window of the 6 preceding and 6 following tokens was used for all features.

As noted previously, data entry staff were able to quickly spot named entities of interest solely because of their visual characteristics. To account for such visual characteristics, we also included visual features associated with text chunks. We used the computed HTML style attributes for each DOM element containing text. Table 5 lists the computed visual features and shows details on how we normalized the style attribute values for font size, RGB color, and Y coordinate.

We then applied SVM on the NER task using the LibSVM library. We again chose SVMs as they have been shown to perform well on a variety of NER tasks, for example (Isozaki and Kazawa, 2002; Takeuchi and Collier, 2002; Mayfield et al., 2003; Ekbal and Bandyopadhyay, 2008). We used a linear kernel model with the default parameters. The multi-class problem was converted to binary problems using the one-vs-others scheme.

Feature Name	Description
Font Size	The computed *font-size* attribute of the surrounding HTML DOM element, normalized to 7 basic sizes (*xx-small, x-small, small, medium, large, x-large, xx-large*).
Color	The computed *color* attribute of the surrounding HTML DOM element. The RGB values were normalized to a set of 100 basic colors. We converted the RGB values to the YUV color space, and then used Euclidian distance to find the most similar basic color approximating human perception.
Y Coordinate	The computed *top* attribute of the surrounding HTML DOM element, i.e. the y-coordinate in pixels. The pixel locations was normalized to 150 pixel increments (roughly 1/5th of the visible screen for the most common screen resolution.)

Table 5: List of visual features used for the NER task. A sliding window of 6 preceding and 6 following DOM elements were used for all features.

As described earlier, we used a portion of the total training data (a total of 1,361 flyers) for the NER tasks. We were able to automatically annotate and use as training data 878 flyers used for address named entities, 1,145 flyers used for broker information named entities, and 1,242 flyers for space named entities. Results were evaluated against the manually entered data for the full test set of 908 flyers. We first used the trained classifiers to find named entities, including their boundaries and types. The predicted named entities were then used to generate listing data as follows. For attributes that have a single value per flyer, we used the predicted named entity of the corresponding type with the highest probability estimate[12]. Single value listing attributes are the fields of the listing address (street or intersection, city, state, zip). Flyers contain a single list-

[12] We used the LibSVM probability estimates for each predicted named entity.

ing, which in turn has a single address. In contrast, broker information and space information are multi-value attributes. A listing is typically represented by multiple brokers and can contain multiple spaces. To construct listing information in the case of multi-value attributes, we used all predicted named entities of the corresponding types. The predicted listing information was then compared to the gold standard of manually entered listing data.

The construction of listing data (for comparison with manually entered data) resulted in a strict performance measure. We consider an answer to be correct only if both the entity boundaries and entity type are accurately predicted. In addition, in the case of single value attributes, only the highest ranking named entity (based on estimated probabilities) was retained.

Results are shown in Table 6. We compared performance of classifiers using only textual features (first 3 columns), versus performance using both textual and visual features (next 3 columns).

Named Entity	Pt	Rt	Ft	Pv+t	Rv+t	Fv+t
Broker Name	93.3	81.2	86.9	95.9	85.5	90.4
Broker Email	95.6	83.6	89.2	95.8	86.5	90.9
Broker Phone	95.4	82.6	88.6	95.7	83.3	89.1
Company Name	97.6	93.9	95.7	98.2	94.9	96.5
Street	77.0	83.4	80.1	81.4	88.6	84.9
City	88.1	96.1	91.9	92.0	98.3	95.0
State	93.1	98.6	95.8	95.4	99.4	97.3
Zip	92.0	86.7	89.3	96.3	86.4	91.1
Space Size	76.8	57.9	66.0	80.1	65.7	72.2
Space Type	66.7	62.6	64.6	68.8	66.7	67.8
OVERALL	87.7	80.3	83.8	89.7	83.5	86.5

Table 6: Results from applying SVM using the textual features described in Table 4, as well as both the textual and visual features described in Tables 4 and 5. t=textual features only, v+t=visual + textual features, P=Precision, R=Recall, F=F1-score

The addition of visual features significantly[13] increased the overall F1-score from 83.8 to 86.5%. Performance gains are more significant for named entities that are typically visually salient and are otherwise difficult (or impossible) to identify in a text-only version of the flyers. In particular, improvements were most significant for named entities referring to space information. A flyer typically describes multiple spaces, however, only a few of these are considered relevant for the purposes of listing services. For example, the size of an office space is typically entered, while the size of an office associated with a retail or industrial space is typically omitted. Similarly, lot size is included in building and land listings, but excluded when the listing refers to a unit (or multiple units) within a building. Essential space information is usually prominently displayed and as a result easy to identify. Similarly, named entities referring to address information also showed overall significant improvement. As noted earlier, the property address (vs other addresses in the flyer) is typically visually prominent. In both cases, visual features proved useful predictors.

4 Discussion

In both the text categorization and NER tasks performance improved significantly over the baseline with the addition of typographic and visual features. However, in both cases, improvements were somewhat moderate (around 3% on average). Further improvement could be achieved by including features that account for additional visual characteristics, such as a measure of how eye-catching or striking the relative font color differences are, the perceived contrast between foreground and background colors, etc.

In future work, we could also add to the overall system an image classification component. It has been noted that occasionally the only indicator of the property type of a flyer is present in embedded flyer images and not present in the flyer text. For example, a number of flyers display images of the inside and outside of restaurants, gas stations, shopping malls and thus specify the property type as *retail* without giving additional textual clues. Similarly, an image of a warehouse, a land parcel, or an areal photo of a shopping center explicitly identify the listing property type.

Lastly, it should be noted that an overall system performance baseline is one that measures the average performance of data entry staff in commercial real estate listing services. However, the terms and conditions of most listing services prohibit gathering and using data for such purposes. We were able to collect a very small set of listings (100 listings)

[13] The difference is statistically significant with p value $<$ 0.05% using Z-test on two proportions.

from several listing services[14] and evaluate the precision of a limited set of listing fields. We compared the values of manually entered listing fields against the associated flyer (considered to be the gold standard). The precision of property type, transaction type, space type, and space size was measured as 97%, 79%, 72%, and 73% respectively. While results are not conclusive, this preliminary evaluation suggests that machine learning could achieve performance on par with the performance of manual data entry.

5 Related Work

A number of studies survey and compare term weighting schemes and feature selection methods for text categorization, for example (Salton and Buckley, 1988; Yang and Pedersen, 1997; Debole and Sebastiani, 2004; Soucy and Mineau, 2005; Lan et al., 2005; Lan et al., 2009). They describe supervised and traditional term weighting schemes. All, however, are only considering the textual information in documents such as the term frequency, the collection frequency, combined with normalization factors, various information theory functions and statistics metrics.

A number of term weighting schemes have been suggested for web retrieval and classification that rely on the HTML DOM structure. (Cutler et al., 1997; Cutler et al., 1999; Riboni, 2002; Kwon and Lee, 2000). The idea is that terms appearing in different HTML elements of a document may have different significance in identifying the document (e.g. terms in HTML titles and headings vs HTML body). In our dataset, however, visually salient information does not fall into any distinctive HTML element type. Instead all text is typically presented in *div* elements whose style characteristics are defined by a number of css descriptors complicated by external css files, css inlining, style inheritance, and browser defaults.

Nadeau and Satoshi (2007) present a survey of NER and describe the feature space of NER research. While they mention multi-media NER in the context of video/text processing, all described features/approaches focus only on textual representation.

The literature on Information Extraction from HTML resources is dominated by various approaches based on wrapper induction (Kushmerick, 1997; Kushmerick, 2000). Wrapper inductions rely on common HTML structure (based on the HTML DOM) and formatting features to extract structured information from similarly formatted HTML pages. This approach, however, is not applicable to the genres of marketing materials (PDF and HTML) since they typically do not share any common structure that can be used to identify relevant named entities. Laender et al. (2002) present a survey of data extraction techniques and tools from structured or semi-structured web resources.

Cai et al. (2003) present a vision-based segmentation algorithm of web pages that uses HTML layout features and attempts to partition the page at the semantic level. In (Burget and Rudolfova, 2009) authors propose web-page block classification based on visual features. Yang and Zhang (2001) build a content tree of HTML documents based on *visual consistency* inferred semantics. Burget (2007) proposes a layout based information extraction from HTML documents and states that this visual approach is more robust than traditional DOM-based methods.

Changuel et al.(2009a) describe a system for automatically extracting author information from webpages. They use spatial information based on the depth of the text node in the HTML DOM tree. In (Changuel et al., 2009b) and (Hu et al., 2006), the authors proposed a machine learning method for title extraction and utilize format information such as font size, position, and font weight. In (Zhu et al., 2007) authors use layout information based on font size and weight for NER for automated expense reimbursement.

None of the above studies, however, include computed HTML style attributes (as seen in browsers), and as a result are not applicable to the vast majority of web pages which do not rely on HTML layout tags or DOM-structure to describe style.

6 Conclusion

In this study, we generated dataset and features from available commercial real estate flyers and associ-

[14] Due to data usage restrictions we were unable to collect a larger dataset or reveal the identity of the source listing services.

ated manually entered listing data. This approach precludes the need for manual linguistic annotation and instead relies on existing data available from commercial real estate listing services. We modeled the structured data extraction task as text categorization and NER tasks and applied machine learning (SVM) on the automatically generated training datasets. The learned models were then applied on our test set and the predicted values were used to reconstruct listing data matching the manually entered fields. Results suggest that this completely automated approach could substitute or enhance the existing manual data entry workflows.

In addition, we have shown that ubiquitous online formats such as PDF and HTML often exploit the interaction of textual and visual elements. Specifically, in the marketing domain, information is often augmented or conveyed by non-textual features such as positioning, font size, color, and images. We explored the use of novel features capturing the visual characteristics of marketing flyers. Results show that the addition of visual features improved overall performance significantly in the context of text categorization and NER.

References

Radek Burget and Ivana Rudolfova. 2009. Web page element classification based on visual features. In *Intelligent Information and Database Systems, 2009. ACIIDS 2009. First Asian Conference on*, pages 67–72. IEEE.

Radek Burget. 2007. Layout based information extraction from html documents. In *Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on*, volume 2, pages 624–628. IEEE.

Deng Cai, Shipeng Yu, Ji-Rong Wen, and Wei-Ying Ma. 2003. Extracting content structure for web pages based on visual representation. In *Web Technologies and Applications*, pages 406–417. Springer.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Sahar Changuel, Nicolas Labroche, and Bernadette Bouchon-Meunier. 2009a. Automatic web pages author extraction. In *Flexible Query Answering Systems*, pages 300–311. Springer.

Sahar Changuel, Nicolas Labroche, and Bernadette Bouchon-Meunier. 2009b. A general learning method for automatic title extraction from html pages. In *Machine Learning and Data Mining in Pattern Recognition*, pages 704–718. Springer.

Michal Cutler, Yungming Shih, and Weiyi Meng. 1997. Using the structure of html documents to improve retrieval. In *USENIX Symposium on Internet Technologies and Systems*, pages 241–252.

Michal Cutler, Hongou Deng, SS Maniccam, and Weiyi Meng. 1999. A new study on using html structures to improve retrieval. In *Tools with Artificial Intelligence, 1999. Proceedings. 11th IEEE International Conference on*, pages 406–409. IEEE.

Franca Debole and Fabrizio Sebastiani. 2004. Supervised term weighting for automated text categorization. In *Text mining and its applications*, pages 81–97. Springer.

Asif Ekbal and Sivaji Bandyopadhyay. 2008. Named entity recognition using support vector machine: A language independent approach. *International Journal of Computer Systems Science & Engineering*, 4(2).

Yunhua Hu, Hang Li, Yunbo Cao, Li Teng, Dmitriy Meyerzon, and Qinghua Zheng. 2006. Automatic extraction of titles from general documents using machine learning. *Information processing & management*, 42(5):1276–1293.

Hideki Isozaki and Hideto Kazawa. 2002. Efficient support vector classifiers for named entity recognition. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, pages 1–7. Association for Computational Linguistics.

Thorsten Joachims. 1998. *Text categorization with support vector machines: Learning with many relevant features*. Springer.

Nicholas Kushmerick. 1997. *Wrapper induction for information extraction*. Ph.D. thesis, University of Washington.

Nicholas Kushmerick. 2000. Wrapper induction: Efficiency and expressiveness. *Artificial Intelligence*, 118(1):15–68.

Oh-Woog Kwon and Jong-Hyeok Lee. 2000. Web page classification based on k-nearest neighbor approach. In *Proceedings of the fifth international workshop on on Information retrieval with Asian languages*, pages 9–15. ACM.

Alberto HF Laender, Berthier A Ribeiro-Neto, Altigran S da Silva, and Juliana S Teixeira. 2002. A brief survey of web data extraction tools. *ACM Sigmod Record*, 31(2):84–93.

Man Lan, Chew-Lim Tan, Hwee-Boon Low, and Sam-Yuan Sung. 2005. A comprehensive comparative study on term weighting schemes for text categorization with support vector machines. In *Special interest* *tracks and posters of the 14th international conference on World Wide Web*, pages 1032–1033. ACM.

Man Lan, Chew Lim Tan, Jian Su, and Yue Lu. 2009. Supervised and traditional term weighting methods for automatic text categorization. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(4):721–735.

James Mayfield, Paul McNamee, and Christine Piatko. 2003. Named entity recognition using hundreds of thousands of features. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 184–187. Association for Computational Linguistics.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

Daniele Riboni. 2002. *Feature selection for web page classification*. na.

Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523.

Fabrizio Sebastiani. 2002. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47.

Pascal Soucy and Guy W Mineau. 2005. Beyond tfidf weighting for text categorization in the vector space model. In *IJCAI*, volume 5, pages 1130–1135.

Koichi Takeuchi and Nigel Collier. 2002. Use of support vector machines in extended named entity recognition. In *proceedings of the 6th conference on Natural language learning-Volume 20*, pages 1–7. Association for Computational Linguistics.

Vladimir Vapnik. 2000. *The nature of statistical learning theory*. springer.

Yiming Yang and Xin Liu. 1999. A re-examination of text categorization methods. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 42–49. ACM.

Yiming Yang and Jan O Pedersen. 1997. A comparative study on feature selection in text categorization. In *ICML*, volume 97, pages 412–420.

Yudong Yang and HongJiang Zhang. 2001. Html page analysis based on visual cues. In *Document Analysis and Recognition, 2001. Proceedings. Sixth International Conference on*, pages 859–864. IEEE.

Guangyu Zhu, Timothy J Bethea, and Vikas Krishna. 2007. Extracting relevant named entities for automated expense reimbursement. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1004–1012. ACM.