Predicting Day-Zero Review Ratings: A Social Web Mining Approach

Thesis

Presented in Partial Fulfillment of the Requirements for the Degree
Master of Science
in the Graduate School of The Ohio State University

By
Zubin John

Graduate Program in Computer Science & Engineering
The Ohio State University
2015

Thesis Committee:
Professor Alan Ritter, Advisor
Professor Eric Fosler-Lussier
Abstract

Social Web Mining: is a term closely associated with modern day use of the Internet; with large Internet companies Google, Apple, IBM moving towards integration of intelligence into their product eco-system, a large number of different applications have popped up in the Social sphere. With the aid of machine learning techniques there is no dearth of learning that is possible from endless streams of user-generated content. One of the tasks in this domain that has seen relatively less research is the task of predicting review scores prospectively i.e. prior to the release of the entity - a movie, electronic product, game or book in question. It is easy to locate this chatter on social streams such as Twitter; what’s difficulty is extracting relevant information and facts about these entities and even more - the task of predicting these Day-Zero review rating scores which provide insightful information about these products, prior to their release. In this thesis, we propose just such a framework - a setup capable of extracting facts about reviewable entities. Populating a list of potential objects for a year, we follow an approach similar to boot-strapping in order to learn relevant facts about these prospective entities, all geared towards the task of learning to predict scores in a machine learning setting. Towards the end-goal of predicting review scores for potential products - our system supports alternative strategies which perform competitively on the task problem. All the predictions from the learning framework, within a certain allowable error margin output scores comparable to human judgment. The results bode well for potential large-scale predictive tasks on real-time data streams; in addition this framework proposes alternative feature spaces which in aggregation go on to describe a multi-method approach to achieving higher accuracy on tasks which have previously seen lack-luster results.
Acknowledgements

I would like to thank and acknowledge in order; Professor Alan Ritter for his support over the entire period of work on this thesis; Professor Eric Fossler-Lussier for his initial comments, thoughts and input to the idea; the CSE department at Ohio State along with the numerous associated people for their support, guidance and prompt administration. Finally, my colleagues and the countless discussions they made possible.
# Table of Contents

Abstract ........................................................................................................ ii

Acknowledgements ....................................................................................... iii

List of Figures ............................................................................................... vi

List of Tables ................................................................................................ vii

1 Introduction ............................................................................................... 1

2 Related Work ............................................................................................. 6

3 Relevant Background .................................................................................. 10
   3.1 Web 2.0 and User-Generated Content .................................................. 10
   3.2 Twitter .................................................................................................. 12
   3.3 Relevant Theory ................................................................................... 14
      3.3.1 Parts-of-speech (POS) Tagging ..................................................... 15
      3.3.2 Support Vector Machines ............................................................. 16
   3.4 Summary .............................................................................................. 18

4 Experimental Setup .................................................................................... 19
   4.1 Dataset Compilation .............................................................................. 19
      4.1.1 Review Objects and Facts ............................................................ 20
      4.1.2 Twitter Dataset ........................................................................... 25
   4.2 Feature Engineering .............................................................................. 25
      4.2.1 Bag of words ............................................................................... 26
      4.2.2 TF-IDF ....................................................................................... 26
      4.2.3 Parts-of-Speech (POS) Action-Relation Descriptors .................. 26
   4.3 Experimental Design ............................................................................ 27
   4.4 Summary .............................................................................................. 32

5 Results and Discussion ............................................................................... 33
   5.1 Preliminary Tests .................................................................................. 33
   5.2 Evaluation Scheme ............................................................................... 34
   5.3 Outcome ............................................................................................... 36
      5.3.1 Results ......................................................................................... 36
      5.3.2 Observable Patterns .................................................................... 39
List of Figures

3.1 Maximum margin classifiers for decision surfaces with varying margin sizes. 17

4.1 Wikipedia page for Master List of all movies released in 2011; we scraped this page to generate a database for our movies 22

4.2 Wikipedia games released in 2011; we filtered results extending up from 12 April 2011 - 11 May 2012 22

4.3 Screenshots of the extraction process (Source: Engadget) depicting scope of information scraped from Engadget page for each product. 23

4.4 Histogram showing distribution of True values of review rating scores across genres (Products, movies and video games - left to right). These values have been used for training our model as well as testing in our evaluation scheme. 25

4.5 Diagram for skeleton of framework and flow of control 29
List of Tables

4.1 Table describing number of entities extracted per each class; for our knowledge base. .................................................. 21
4.2 Snapshot of Products fact table, consisting of product name, release date and truth value (Engadget product score, in this instance) associated with the product .................................................. 24
4.3 Snapshot of Tweet’s associated with IPhone 4S exemplifying the contextual information in associated entities. ............. 31
5.1 Results of initial training and testing on a probe sample data-set of 100,000 tweets. Different regression methods have been summarized with the best results from each. .................................................. 34
5.2 RMSE and coeff. of determination for each model in our research: automated predictors evaluated against 5-fold cross validation with results reported for highest accuracy across all folds. Different feature models and learning algorithms have been compared. .................................................. 37
5.3 Table describing useful statistics for each class in our knowledge-base. . . 38
5.4 Highest regression coefficient for each review entity class along with corresponding feature. .................................................. 40
Chapter 1

Introduction

The growth of opinion mining and sentiment analysis over the past decade has been remarkable; its development molded in large by the recent revolution in Web 2.0 generated content and social media. (Pang and Lee, 2008) sparked interest within the community, on the need to develop sentiment aware intelligent agents and its potential. This lead to a renewed interest in re-evaluating a number of existing systems with the assumption that underlying sentiment, along with structure, should be taken in to account towards predictive analytics. Despite growing popularity, predicting any form of review ratings is a domain problem that has yet to profit from the burgeoning attention.

The growing popularity of online product review forums motivates the need for agents capable of analyzing and making sense of this data. There is certainly a burgeoning interest in products and the study of their usefulness. Reviews offer much needed context to a product while skimming over details; trading verbosity for efficiency. As such, agents capable of not only generating informative decision parameters for products, but doing this prior to product release - would be a valuable market commodity. Currently, human experts are well enough suited to the task - for instance Cinemablend\(^1\) a popular movie-related entertainment portal does a regular section on trying to predict the *Tomatometer\(^2\)* ratings for upcoming movies on a regular basis. Like most artificial intelligence tasks, human experts given a suitable amount of domain knowledge could beat any intelligent agent at this task.

\(^1\)[http://www.cinemablend.com]
\(^2\)The Tomatometer is a metric used by popular movie rating site [http://www.rottentomatoes.com] which rates movies on a percentage scale indicating ‘freshness’ or alternatively ‘rotten-ness’.
hands down. However, here again sheer volume and attainability are the regular challenges faced by interested parties. Interested parties - marketing and advertisement agencies, have difficulties often getting experts into their strategy teams to navigate the vast ocean of information required to draw significant conclusions. In the past review related tasks have been rather unambitious in aiming simply to provide a more indicative review rating metrics, fueled by the argument that aggregated scores on communities such as Amazon can often be misleading and ‘unhelpful’.

Traditional review rating systems as discussed in (Qu et al., 2010; Ganu et al., 2009) target this problem of getting more meaningful rating scores from textual features of the review itself. While, this is no-doubt a strategic tool to marketing and product support teams; it does not respond well to the challenge of a more prescient form of predicting the reviews as followed by human experts. We believe it is quite possible to perform similar tasks using machine learning to obtain what can be easily termed as ‘Day-Zero’ reviews. This is made possible by analyzing common keywords and lexical trends that associate a potential review entity to historic reviews of similar objects. Toward this purpose we seek to propose an approach that discards textual review content for a novel means of extracting relevant information corresponding to our review entities.

Perhaps key to any discussion regarding predictive analytics is the side-effect of evolution of social media. Social media of course, provides this much needed stream of alternate content; predating the actual review text for performing these predictions. It has been the link to exploiting pluggable data, a steady feed of content that has enabled task-systems in sentiment analysis, event extraction and box-office revenue interpolation; a new avenue in applications. The work described in this thesis, seeks to build upon this available platform - provide a more potent tool towards drawing early insights by providing review rating scores.

³http://www.amazon.com
For better of worse, we have decided to proceed in this direction without any reliance on the need to exploit the underlying sentiment of our review entities. What we have succeeded at is building a highly generic intelligent agent, capable of making these Day-Zero predictions with limited training on data similar to that which it is most likely to see in real-time application. Our results, indicate a measurable amount of success in predicting rating scores for movies, high-profile consumer electronics and video games, prior to release.

To our best knowledge, there is no other work that relies solely on textual features to predict these rating scores for a review entity. In similar work (Oghina et al., 2012) use surface features boosted by filtered textual features in a cross-channel prediction task via looking for signals in Twitter and YouTube to predict ratings for movies on IMDB. Our task, differs in its rather generic approach to Day-Zero predictions for any class of review entities with a primary focus on textual features. It follows an information retrieval styled approach to extracting relevant textual trends and progressively expanding this set for each review entity. In the grand scheme, it is a pluggable agent which transforms easily to a deployable, productive system capable of analyzing real-time streams and utilizing regression to score objects. As a consequence, our experimental results may perhaps be sophomoric but are highly indicative of a good direction for future work. It provides impactful insights into the feature engineering related to the predictive analytics task when following a social web mining approach.

Most review systems use either a 10 point reviewing system or a percentage scale. The ten-point review system can be said to be highly pervasive and within the scope of our work we have chosen to stick with this system. Almost every reviewable entity that has a catered social community or rating web-site follows this system or a variant of it. In accordance to this semi-standard metric we propose a system in the course of this thesis that is capable of making similar review rating predictions on a ten point system. We have
considered movies, video games and consumer electronics products - each of which has a catered service providing review scores on a ten-point band for its subscribed communities. We use these ratings as truth values via web mining techniques to train a system which is capable of making similar predictions for products that have yet to see the market. Hence, we have achieved a significant level of success in the task-problem of prospectively predicting reviews prior to any textual review for the entity being released on one of the dedicated channels. We term these review scores ‘Day-zero’ review-ratings following the ideology that these reviews are actually available as a direct consequence of our system, prior to actual mass-scale release.

Another matter of import in our work is addressing the growing sensitivities around the ethics of data mining. With the concern of data privacy and security in the Internet community; Internet-based companies are highly reluctant to trade personal information regarding their customers and users. If this regular source of data was exhausted a lot of the predictive analytical systems would soon face obsoleteness. A transient solution to this, is offered in the form of knowledge mining systems, which serve to be the next step in the evolution of data. With this approach systems attempt to capture more ‘knowledge’ rather than actual stored information, drawing conclusion through processing publicly available information such as social media. It is towards this objective that we envisioned a system which would rely on the knowledge paradigm rather than the information paradigm. It is in this concordance, we concluded that a system based on textual analysis of public feeds would be more profitable than one relying on features from multiple channels.

The way this document is organized, Chapter 2 covers a brief overview of related work in predictive analytics and social web mining. We draw attention to recent trends in this area of work which indicate the feasibility of predictive tasks in social media. Chapter 3 covers a lot of groundwork and theory which makes a lot of our work possible, we provide
a brief round-up of the social media era and mathematical foundations of algorithms and machine learning approaches that we have used in our work. With Chapter 4, we start diving in to the bulk of our work, describe the feature engineering task which set’s up the preliminaries on data mining and information extraction for our research. We follow this up with a description of the experimental framework that needs to be setup in order to collect our observations. Chapter 5 describes our results, and the highlights some of our findings and discussion points moving forward. Finally, in Chapter 6 we round-up with a conclusion summarizing the scope of our work and leaving with implication for future research in this track.
Chapter 2

Related Work

Social Media has played a major role in shaping the landscape of the research surrounding Natural Language Processing, information extraction, opinion mining and review prediction. (Pang and Lee, 2008) was perhaps iconic in its time, tracking recent developments in sentiment analysis laying the foundation for subsequent research to follow. With the evolving social-media interest it wasn’t long before researchers forayed into experimenting with intelligent agents which performed opinion mining on social media feeds as well. Opinion mining received a lot of interest from the scientific community - branching into sub-tasks of sentiment classification, subjectivity classification, text summarization and topic modeling all under the all-encompassing umbrella of studying user opinion. Concurrently, vast applications of sentiment analysis and the burgeoning social media scene were also being explored by the scientific community. (O’Connor et al., 2010) drew parallels between textual analysis of sentiment and their correlation in public poll opinion, in other work (Golbeck et al., 2011) worked on predicting user personalities on social media web-sites.

The stage was set for tasks predicting quantitative factors through analysis of social media data. (Bollen et al., 2010) was of particular interest as they reported relative success in the task of predicting stock market indicators by incorporating sentiment. In similar work, (Zhang et al., 2011) extended on previous work to demonstrate a correlation between daily changing volume in economic trends and value shift in market indicators the next day.
With Machine Learning and specialized classification models leading the forefront in prediction; tasks such as estimating review scores and movie ratings were soon possible. Mining text reviews was a task that had been explored previously with (Hu and Liu, 2004) but saw a resurgence with the social media era. Pre social media there had been some experiments with predicting the usefulness of reviews, (Liu et al., 2008) looked at predicting helpfulness of reviews via an empirical study of IMDB reviews; (Zhang and Zhang, 2014) in more recent work revisits this task by predicting helpfulness of reviews using an SVM model trained on Amazon reviews.

Predicting actual numerical scores for reviews on the other hand, was a task that had seen relatively lesser success, (Qu et al., 2010) proposed a novel bag-of-opinions approach to overcome the limitations of the preexisting bag-of-words and bag-of-phrases approaches by introducing a ridge regression based model trained on review scores for each sample. (Li et al., 2011) offered fresh insight to this task with the argument that predicting numeric review ratings was related not only to the text of the review itself but features of the reviewers or the authors as well. Both these proposed frameworks resulted in better performance by overcoming the data sparsity bottleneck.

Meanwhile in the social media space, (Asur and Huberman, 2010) studied pre-release information on social media feeds to train and predict box office revenue for movies, prior to their actual release. This signified a transition; social media had finally captured the attention of a scientific majority - it addressed the challenge of information and generation and offered a new avenue for applications. In the coming years, the research community saw many attempts at exploiting this source of knowledge and put it to a myriad applications. Perhaps a precursor to work presented in this thesis, (Oghina et al., 2012) use surface features boosted by filtered textual features in a cross-channel prediction task, via looking for signals in Twitter and YouTube to predict ratings for movies on IMDB. (Lica and Tuta,
2011), follow an alternate approach of correlating sentiment indicators in the Twitter search space to predict product performance.

Predicting numerical quantifiers from Twitter is just part of a largely interesting problem that addresses the question - How feasible is it to make predictions regarding events yet to happen purely from large streams of short-text messages? Lots of modern research asks innovative questions and using machine learning engages in the task of finding answers to these questions. Within this domain of exploring predictive capabilities of Twitter, in recent times we see varied research problems of interest. (Chang et al., 2012) explore the problem of predicting home location of Twitter users using a blend of three different probabilistic approaches. Other work in the trending domain of Computational Social Sciences aspire to bring computational solutions to interesting questions in the field of Social Science. Politics is a sub-domain of remarkable interest in this field and we see different problems being tackled through intelligent agents in the Twitter-sphere. In their work (O’Connor et al., 2012) link public sentiment analysis to mirror polling results in an approach that models user opinions through correlating word frequencies in Twitter streams. In similar context (Tumasjan et al., 2010) explore the very same problem of predicting election results through an approach that is inclined towards analysis of psycho-linguistic cues in Twitter streams.

A lot of the recent work in the social media domain, indicates an upward trend towards asking questions of the nature - can we make future predictions with data in public social streams? This is certainly a strong motivation behind our work, to test the theory on how twitter streams can contribute towards extracting signals capable of indicating numerical quantifiers for future events. The event of interest in our work being - release of movies, electronic consumer products and video games in the future; and subsequently predicting how well these entities will be reviewed (estimate a score) well in advance.
Our work in this thesis, aims to combine some of the recent approaches cited in this chapter, to predict review ratings in a prescient manner. An important objective for our work was to be able to produce results which would define a generic model capable of predicting quantitative scores for entities across genres. Towards this goal, we have focused more on textual features rather than surface properties and cross-channel features, which would limit design-flexibility of the system. As a result our framework follows a pluggable data model, demonstrated on three different genres of product categories - consumer electronics, movies and video games. While we achieve different performance measures across the data-sets, we are able to define a unifying framework for the necessary feature engineering and a model for predicting review scores across heterogeneous review parameters.

Much of the feature spaces described in our work are based on modern parts-of-speech taggers that are built especially for the informal, unstructured datasets found on social media channels such as Twitter. (Gimpel et al., 2011) present a Twitter part-of-speech tagger that performs with high accuracy on their annotated data. Similarly (Ritter et al., 2011) attempted to perform Named-Entity Recognition and made available their tagger to the research community. The availability of these Twitter NLP tools have enabled a vast amount of research analyzing social media text towards downstream applications. Our work relies heavily on these tools and goes on to show how, working with these feature spaces can make for higher precision intelligent agents in downstream predictive tasks. In successive chapters we will demonstrate our use of these tools to engineer and design experiments for predicting from social media streams independent of opinion and cross-channel signals.
Chapter 3

Relevant Background

Milestones in the development of the World Wide Web have always been major disruptor’s to how the audience perceives technology. The Internet in all its metamorphic glory transcended to a higher state, albeit unnoticed by many with the introduction of Web 2.0. For a good decade or two, the relationship between the Internet and its users had been primarily one-sided. News, information, content and even entertainment was catered to the audience with little feedback; the first Web 2.0 conference hosted by O’Reilly Media \(^1\) was an indication of just how drastically this relationship was soon about to change.

3.1 Web 2.0 and User-Generated Content

The definition of Web 2.0 is perhaps as elusive as that of the Internet; however as of today, the Web 2.0 signifies the next phase or the second generation of the Internet itself. As emphasized, the previous iteration of our Internet relationship spoke highly of one-sidedness; this was forever changed when the Internet conglomerate decided to put content in the hands of users as well. Information could now be propagated not only as a traditional subscriber based service, but each subscriber in turn would be responsible for content creation. This could be done in the form of blogs - easy to edit personalized micro web-sites, forums which offered platforms for users to ask questions an in-turn respond and finally wikis which

\(^1\)O’Reilly Media ([www.oreilly.com](http://www.oreilly.com)) is an American media company founded by Tim O’Relly. Its primary services include publishing books and web-sites relevant to computer technology. On occasion has organized conferences in the past.
were large user-contributed and moderated encyclopedias of worldly knowledge.

This advent, of course was not devoid of its own pros and cons. With user-generated content came mass participation and user involvement; the ability to have free user-activated classification of information via tagging and collective collaboration, and finally content that is dynamic and shaped by human stimuli. On the flip-side with new content generated every day without restrictions the Internet space erupted with an explosion of content. Content: in the form of images, videos, test and audio. With the sheer scale of data coming in at times, it is difficult to segregate signal from the noise. Users still have the ability to subscribe to media of their preference, however with no curation in place, finding relevant content can sometimes be as challenging as navigating a tumultuous ocean. This need for organization and better indexing of information lead to a demand for intelligent agents. Agents capable of understanding users need, making intelligent searches using historical and personal context, the ability to analyze images and to recognize voice commands. This is the next generation of a smarter Internet age - content and information broad-casted and generated almost ubiquitously, it has also led to the introduction of a whole new class of devices to enter the arena competing for the user’s attention.

As mobile devices entered the foray, so did a new segment of data generation capabilities. Mobile devices engaged a wider audience and drove Internet traffic to new heights. Within a few years, mobile devices were soon responsible for majority of all online traffic and a significant fraction of online content. Image and video content creation climbed exponentially with mobile applications that allowed users to send images and video without having to sync via cables to a computing device. Audio messaging and telephony also saw great rise in usage but never quite caught up to the hype of their counterparts.

Perhaps more of interest to the scientific community was the creation and generation of
content that could be studied and analyzed; a source that was capable of better bridging the gap between the vast ocean of signals and noise. The unlikely candidate to bridge this gap was soon found in social media. By definition a social website is a portal for users to connect with their email contacts, make new acquaintances and discover friends; all the while creating a graph of connections that eventually would be termed as a network. With the growing need for individuals to be social with their online presence; users would periodically indulge in sharing experiences within their network via a cacophony of media. This media would then offer a quantifiable footprint of users on the Web, an identity of sorts - siphoned through a vast collection of their shared thoughts, images and recorded recollections. Here was a potential for vast information; true knowledge bases of human thought, emotion and opinion. Social media through its clustered nature offered signals that otherwise required years to capture or generate. It is this aspect of the nature of social media that attracts the attention of the academic community. In a world where mixed information is prevalent, social networks have shown themselves to be a highly accurate mirror of the general publics’ perceptions.

### 3.2 Twitter

Early 2006, saw the launch of a new online platform, **Twitter** was initially perceived to be just another online "social networking" platform in the footsteps of Orkut\(^2\) and MySpace\(^3\). While the latter two never really made more than an initial splash - Twitter on the other

---

\(^2\)Orkut was perhaps considered the precursor to social media, established in 2004, it could easily have ridden the social media wave - however a lack of vision and slow growth in the face of competition, lead to its demise.

\(^3\)Another pioneering social web-site MySpace was considered a big player in its earlier days; however the advent of Facebook and Twitter took away a large portion of its fan-base. MySpace was soon second seed to the tech-powered approach of the duo; it has only managed to remain in existence due to its pivot into a music platform to connect audience and artists.
hand literally changed the road map of social media and online networks. Facebook\(^4\) and Twitter, the big two of social media; grew to dominate this sector of the technology market in the coming years. While Facebook may have taken the lead in the race to capturing a larger user-base, in recent years, Twitter seems to have caught up to the competition. Twitter - though a social media website; follows quite a different model; users “Tweet” a unique form of micro-blogged messages that may depict their thought, mood or a matter of opinion to their subscribed followers. Each Tweet is limited to 146 characters and in a naive sense, as most social media researcher have come to believe, contains more signals when compared to more verbose forms of text communication - when correctly harnessed off-course! Coerced by competition, Twitter evolved as required and today has incorporated more multimedia elements. Its modern day offerings include links, images and embeddable short videos known as “vines”.

Twitter was the first social media site to proponent the use of hash-tags (or #hashtags). Hash-tags allowed the clustering of similar topics or Tweets under a single index; popular topics of discussion would hence catch the eye of the public and in-turn generate more content through continuous references. Hash-tags in turn resulted in widespread popularity of certain topics around the globe which were termed to be trending. In one stroke Twitter was now an effective mirror for reporting, propagating news and information for the worldwide online community.

Hash-tags and trends provided a means to link online media and content to real world events. Pretty soon Twitter overtook traditional journalism as a source for real-time feeds and media for concurrently occurring global phenomena. It is often argued that Twitter succeeds in capturing certain information that is beyond the scope of traditional sources;\(^4\)

\(^4\)The largest online social platform and the most visited web-site in The Unites States - needs no introduction.
nuances such as general reaction to trending news, feedback and volume of knowledge being propagated corresponding to an event.

These aspects of Twitter have allowed its transformation into a preferred tool and source of knowledge for a large base of researchers from every branch or discipline. Signals on Twitter range from political inclination, to live real-time insight into events, of most interest to us - public feedback on movies, consumer products and even novelty items such as games and books. Patterns in social media chatter for these objects make it possible for us to predict prospective reviews, feedback and reception quantifiers for them. We explore these elements over the course of this thesis in the hopes of uncovering a framework for intelligent agents to analyze large streams of data and learn to predict with minimal training - feedback and other quantifiers for success of potential reviewable objects. The challenge here is two-fold, identification of objects that are potential candidates and then later moving on to the estimation task for feedback quantifiers. The bulk of the work relies on open event extraction of these objects as elaborated in (Ritter et al., 2011). We will not review the background for successful extraction of objects nor the results of the above mentioned system. The thesis will however go on to set up a downstream task and provide relevant background for comprehension of the proposed framework. The following sections go on to elaborate on some core fundamentals that are key to our framework.

### 3.3 Relevant Theory

In this section we review a lot of the prerequisite theory which will most likely be referenced over the course of this thesis. This includes common terminology in the natural language processing domain, along with the mathematical/statistical principles which form the underlying principle of our learning framework.
3.3.1 Parts-of-speech (POS) Tagging

Language has its own structure, rules and grammar; semantic aspects and syntax which form the distinction between coherence and gibberish. This very property of order allows different classes of words to be distinguished and categorized by tense, form and position. This task of systematic parsing of sentences - generating tokens which indicate the words part of speech is known in computational linguistics as parts-of-speech tagging or POS tagging. With evolution of methods in Natural Language Processing, POS tagging has evolved into a crucial preprocessing step for almost any application system. Perhaps one of the reasons for its omnipresence is that modern parts-of-speech taggers are considered a ‘solved problem, this basically means that most systems assume 100% accuracy on tagging.

With this dependence, for traditional NLP tasks (non-noisy newswire text), POS tags are one of the most commonly used feature sets. Several systems in fact rely as heavily on parts-of-speech as on the bag-of-words approach to feature representation. POS taggers have evolved over time, starting out being trained via Hidden Markov Models to their modern counterparts trained on Conditional Random Fields.

With new research on extracting information from Twitter as a data source gaining popularity; it became painfully obvious that popular tools for linguistic analysis on text performed poorly on Tweets. It was not difficult to see why, as a majority of these tools were trained on newswire text which had relatively fewer challenges of noisy, un-normalized text as seen on Twitter. Two of the fundamental linguistic analysis and feature preprocessing tasks have been re-evaluated under the scope of Twitter data, resulting in powerful tools better suited at tagging and chunking tasks in the social media domain. (Gimpel et al., 2011) and (Ritter et al., 2011) contributed to the domain of Natural Language Processing for Twitter with their respective work in developing tagsets and training corpuses for the parts-of-speech tagging and named entity recognition task (NER). The work in this thesis, explores how
POS tags in Twitter contribute to learning discriminatory signals which can be used to train supervised frameworks capable of estimating feedback quantifiers for objects of interest from a real-time stream of Tweets.

### 3.3.2 Support Vector Machines

Support vector machines first introduced by Vapnik(2005) and (Boser et al. 1992) proposed a new kernel-based learning algorithm that was initially applied for classification tasks and later extended to regression as well. The theory behind Linear SVM is that given a set of training examples, a SVM model capable of representing these examples as points in space, strives to learn a linear decision surface capable of distinguishing incoming samples (points) to one side of the decision surface, hence in essence assigning a class to it. The decision is learned uniquely by maximizing a learning margin on either side of the decision surface. By property of learning weights over the training samples the model describes support vectors (the subset) enabling it to achieve capacity to extract globally optimal solutions, irrespective of the sparsity of training features. In addition to performing exceptionally whilst minimizing the adverse effects of training, SVMs also employ the “kernel trick”. The kernel trick allows the SVM to operate in a higher dimensionality space; by projecting non-linearly separable training data onto a higher dimensional space - a linear separation may be achieved. The advantage here is that the kernel method may operate, in a higher space without any actual computation of points in the extended feature space by simply computing inner dot-products.

The general form of this problem is a solver for the optimization given by:

$$\min_{w,b,c} \frac{1}{2} w^T w + C \sum_{i=1}^{i} \xi_i$$

Here $\xi_i$, is the indicator for the degree of errors termed as the slack variable, where

$$y_i (w^T \Phi(X_i) + b) \geq 1 - \xi_i \text{ and } \xi_i > 0$$
Figure 3.1: *Maximum margin classifiers for decision surfaces with varying margin sizes.*

SVMs can also be used for regression with minor adjustments to the classification problem since, put involves estimation of real values - a margin of tolerance ($\Xi$) is described.

$$\xi = \begin{cases} 
|y(x) - \hat{x}| - \xi & \text{if } |y(x) - \hat{x}| \geq \xi \\
0 & \text{otherwise}
\end{cases}$$

For a value $\xi > 0$, this function gives an output of zero if the absolute error between predicted value and tag $\hat{x}$ is less than the value of $\xi$. Hence, again like classification the support vector regression depends on a subset of the training sample which comprise the support vectors; points that are close to the threshold value are ignored. As a result the margin with is described by those training points that lie outside the $\pm \xi$ region around the margin.

The widespread use of support vector machines is attributed to their superior performance in tasks that involve learning over sparse textual features. The SVMs manage this task elegantly without risk of over fitting via dependence only on a subset of training samples
known as the support vectors. An added incentive to training SVM’s is the ability to model even nonlinear relations via use of kernels - this however shifts the problem from one of optimization to one of model selection; which has its own pitfalls. This summary of SVMs is in no way doing justice to the vast amount of theory and possibility behind its applications, a more elaborate tutorial on as the regression can be found in (Scholkopf and Smola, 2003).

3.4 Summary

In this chapter, we introduced the expansion of the Internet into a completely new avenue which offered prospective opportunities for research on a new genre of data. This incoming data had its own set of challenges associated with them, and we hope that we have done a satisfactory job in summarizing these challenges and the respective research tools which were aimed at tackling these challenges. We’ve covered the growth of social media and its role as a game-changer, along with the research that has sprouted up around this particular domain. Finally, we provide a brief description of the learning model we have selected for our task - the support vector machine with a brief description of how this task is framed in both the classification and regression settings. With this chapter wrapping up all prerequisite knowledge, we firmly move on to demonstrating our research and experimental settings next.
This chapter covers the entirety of our experimental set up - the framework for capturing historical data to build our data sets; the feature engineering and processing involved and finally the learning framework which enables the prediction task. The task at hand to predict a review score (on a band of 1 - 10) for reviewable entities (can be done cross domains for movies, video games and products) purely from examining textual features mentioned in respect to the entity on social streams prior to actual entity release. The bulk of this chapter will focus on the Data Gathering task which is of import in addressing a question of how successful prospective predictions can be. We rely in tandem on data coming in from social media; and a knowledge bank listing target entities that are of interest. The latter driving the extraction of relevant information from the former. In addition, we cover in brief the feature engineering involved to make the learning task possible. Finally we cover details of the experimental framework we set-up in order to capture critical results that we expand upon in further sections.

4.1 Dataset Compilation

Bulk of our knowledge base can be broken down into two -

1. Various reviewable entities by genre (movies, games and consumer electronics) with facts and associated attributes which is review scores, release dates.

2. Historical Twitter Dataset A knowledge bank of Tweets following (Ritter et al., 2010) consisting of Tweets between April 2011 and December 2011; containing meta-data
about the Tweets as well as TempEx (Mani and Wilson, 2000) resolved estimations of the Tweets time. Further, all the tweets have been treated to parts-of-speech (POS) tagging before saving each tweet to our data-set.

4.1.1 Review Objects and Facts

For the work detailed in our thesis, we have considered three reviewable object classes of interests namely: games, movies and consumer electronics. We also performed an initial probe study to determine if it would be feasible to estimate review ratings for books and music albums as well, however our study concluded that quality of data would be ill matched for such estimations. For books there were not enough mentions in our data set to satisfactorily resolve the Tweet to be clustered under a particular object entity. An alternative approach would have been to train a rule-based extractor on key-words such as best-seller, however given our initial results, it seemed prudent to avoid integrating these objects into our knowledge base.

Another challenge was the scope of English-language books, far outnumbering objects such as games, movies which led us to believe that estimating book reviews would be a problem better studied in isolation.

In regards to estimating music albums we again faced an extraction challenge - with it being highly onerous to classify with a high confidence if the Tweet simply mentioned news gossip about an artist or actual mention of an album release. Without a strong rule based/machine learning-based approach to this domain specific extraction - any downstream application would have been futile.

For each of the game, movies and consumer electronics domain we were able to set up a fact extraction system. Each class of review objects has a community that caters to
knowledge of these classes. The top-rated knowledge source for games is perhaps undisputed - IGN; IMDb a great source for movie knowledge and consumer electronics being subject to opinion we decided to go with selection of Engadget.

<table>
<thead>
<tr>
<th>Review Class</th>
<th>Extracted Entities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consumer Products</td>
<td>49</td>
</tr>
<tr>
<td>Movies</td>
<td>207</td>
</tr>
<tr>
<td>Games</td>
<td>235</td>
</tr>
</tbody>
</table>

Table 4.1: *Table describing number of entities extracted per each class; for our knowledge base.*

For each of the object classes, we queried reviews for the years of 2010 up to the end of the first quarter of 2011. IGN reviews can be extracted via the Mashape\(^1\) IGN Review API while the IMDb data was queried via the OMDb API\(^2\). With the absence of an API for consumer electronic reviews, we went the circular route of scraping review scores and release dates for entities in our master list. The master list for each category was extracted from the Wikipedia page of all concurrent releases in that year.

\(^1\)www.mashape.co is a collaborative developer based API platform providing developer-friendly APIs for numerous tasks.

\(^2\)Open Movie Database service that programmatically serves data scraped from IMDb
Figure 4.1: Wikipedia page for Master List of all movies released in 2011; we scraped this page to generate a database for our movies

Figure 4.2: Wikipedia games released in 2011; we filtered results extending up from 12 April 2011 - 11 May 2012
The exception - since there is no ordered list of products (consumer electronics) via year indexed on Wikipedia we used Freebase\textsuperscript{3} to extract all electronic products under category ‘computers between a 10 year span of 2000 to the end of 2010’ to generate a list of product lines. We then used co-references of each product line with the results of their corresponding query to Engadget to determine if they were within the date range of our interest before saving them to our knowledge base. In conjunction, for all products that did meet our criteria we also extracted the truth value of the product review rating (later used for training our model and testing against it) from the very same Engadget review page for each of these products. In the case of movies and video games we were able to query the above mentioned API services to get the desired truth value for our review entities.

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{figure4_3.png}
\caption{Screenshots of the extraction process (Source: Engadget) depicting scope of information scraped form Engadget page for each product.}
\end{figure}

As a final step of filtering, we truncated all reviewable objects that lacked a mention in our Twitter data set. To ensure preservation of Tweets with contributing signals, we

\textsuperscript{3}An open source knowledge-base for structured facts
Table 4.2: Snapshot of Products fact table, consisting of product name, release date and truth value (Engadget product score, in this instance) associated with the product.

<table>
<thead>
<tr>
<th>Product Name</th>
<th>Release Date</th>
<th>Engadget Review Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apple iPhone 4s</td>
<td>2011-10-11</td>
<td>8.6</td>
</tr>
<tr>
<td>Sony VAIO F Series</td>
<td>2011-06-21</td>
<td>7.2</td>
</tr>
<tr>
<td>Kobo eReader Touch</td>
<td>2011-06-13</td>
<td>7.5</td>
</tr>
<tr>
<td>Apple iMac 21.5-inch (mid 2011)</td>
<td>2011-05-04</td>
<td>8.2</td>
</tr>
<tr>
<td>Amazon Kindle 4th-gen</td>
<td>2011-09-29</td>
<td>7.8</td>
</tr>
<tr>
<td>Amazon Kindle Fire</td>
<td>2011-11-13</td>
<td>7.1</td>
</tr>
<tr>
<td>Apple Apple TV</td>
<td>2012-03-14</td>
<td>8.5</td>
</tr>
<tr>
<td>Sony PlayStation Portable E-1000</td>
<td>2011-11-30</td>
<td>8.0</td>
</tr>
<tr>
<td>Sharp AQUOS 6-Series</td>
<td>2012-03-13</td>
<td>7.5</td>
</tr>
<tr>
<td>Kobo Vox</td>
<td>2012-05-10</td>
<td>5.4</td>
</tr>
</tbody>
</table>

refrained from eliminating Tweets with reference to a parent brand or company even if the entity itself had not been mentioned. To achieve this we used a Jaccardian similarity metric to analyze mentions named entities against every object in our knowledge base. Thresholded at a satisfactory 0.8; we marked Tweets with an insignificant level of contributing signals for deletion. The final counts of entities for each review class after all steps of filtering, has been recorded in Table 4.1. Meanwhile, Table 4.2 depicts an actual slice of our Products Fact Table - which is used for training and curating our model in a later stage. As reported in Table 4.1, the break-up of entity distributions across genres is not even, however for the most part all three categories follow a Gaussian distribution in terms of truth value scores as depicted in the Histogram plot below.
4.1.2 Twitter Dataset

To evaluate our system we used the data generated for (Ritter et al., 2011) in the open domain event extraction task. To ensure low overhead, we have attempted to minimize the training set for our model while still maintaining a high level of accuracy on our estimations. The data hence used is a subset of 15 million Tweets isolated from a larger set which was collected while crawling Twitter daily over a period of more than two years to generate a large dump of data. For our purpose we have limited the Twitter stream corresponding to product, movie and game launches between 12 April 2011 and 11 December 2011 however the Tweets in this range may reference products released up to 11 May 2012.

4.2 Feature Engineering

Filtering and selection of the right blend of features in text-based machine learning tasks is a recipe for the best results. Our work experiments with some of the most well-known and received feature spaces; comparing results trained on each.
4.2.1 Bag of words

The bag of words approach, perhaps the most common in Natural Language Processing and information retrieval tasks; nevertheless a task that has the potential to return surprising results. Always a start, the bag of words basically represents each document in a set of words. The set - often termed as a vector, disregards order and grammar in exchange for term frequency. The vector is generally a $M \times 1$ dimensional matrix, where $M$ denotes the size of the vocabulary.

4.2.2 TF-IDF

TF-IDF is an extension of the same scheme of bag of words with an added term-weighting which results in a slightly superior feature space model in related tasks (as seen in previous relevant research). TF-IDF is one of those weighting models; TF-IDF short for term frequency-inverse document frequency: which indicates just how statistically significant a term is over the document collection space. The term frequency component is equivalent to the traditional frequency based bag of words scheme. The additional inverse document frequency accounts for just how frequently the term occurs over the entirety of the document space. With this weighing scheme words that occur less frequently are assigned a stronger weight than their counterparts which have a higher frequency (for example the word ‘the’). TF-IDF is computed as a product of these individuals scores:

$$tf-idf(t, d, D) = tf(t, d) \times idf(t, D)$$

4.2.3 Parts-of-Speech (POS) Action-Relation Descriptors

Our final feature model incorporates semantic information using selective POS features. In a short document space verbs, adjectives and nouns can be said to describe the entirety of the document without any loss of substantial information. To test the veracity of this claim, our final model captures only words from the vocabulary (generated for all tokens in the
data-set) that fall under the verb and noun class of POS; to describe actionable relations which are descriptive of events. Since most of the reviews have an associated event that needs to show up on social media stream, it is our hope that these events encode enough integral signals to allow estimations on their basis alone. The truth to this claim, will be reinforced in our results. Adjectives were also considers in our tests but failed to perform above our baseline, whilst increasing complexity of the system - hence were discarded as a feasible approach.

4.3 Experimental Design

Our framework is not incredibly complex in construction and quite contrarily it follows a simple pipeline that makes these predictions possible.

The first stage consists of extraction of knowledge and facts (building our knowledge bank) which serves to reinforce the learning from Twitter streams. In the absence of complete knowledge the database must be set up with objects of interest dictating facts to be retrieved by the extraction framework. Via successful extractions the database of reviewable objects may expanded into a knowledge base of complete information - regarding the entity’s release date and its true rating score. Our review event analyzer dives into this knowledge base to search for the same review entities in the Twitter data set. This process has a search space complexity of \(O(KD)\) for \(K\) reviewable objects over the documents search space \(D\). For each review entity a collector builds a time ordered set of Tweets; with the help of pre-known facts we are able to use a support vector machine to learn relationships between support vectors and ensuing review scores. The system is then capable of predicting review scores for entities which may pop up in social media chatter.

The event extraction based approach follows logically and intuitively for the review classes
of games, movies and products which have large-scale product launches identifiable well enough in advance. Games and consumer products are frequently tested prior to releases; conventions and conferences (WWDC and Google I/O for products; E3, WES for video games) have historically been resolved to calendar events on Twitter chatter- it is only a matter of analyzing and segmenting these clusters to identify mentions of entities of interest. Similarly movies are anticipated weeks before their opening date; with audiences tweeting well before the first review is published-it is quite fruitful to gauge public feedback quantifiers from these sophomoric streams.
Figure 4.5: Diagram for skeleton of framework and flow of control
The first phase of our system is responsible for expansion of knowledge base; this process involves introducing additional knowledge/facts for each known review object. The only information that we know prior to this stage about review objects, is potentially the date of their release. These dates are hypothetically the ones that we would be extracting from our Twitter stream of interest. Once the review objects and their potential dates are known (the dates can be extracted from publicly available knowledge bases such as Wikipedia and FreeBase - this is exactly how we have extracted the basic facts for dates regarding our objects of interest in the course of this work), the next step is to collect facts that would be essential to training of the model. These facts specifically the true review scores are the labels used in the training process; these truth-values been extracted and added to our knowledge base via the use of third party services and APIs, and in dire situations the application of target-specific data scrapers.

At this point where we know a target date and the truth value associated with events; we essentially want to select those tweets as candidates for each entity which most closely provide information relevant to it. We start inductively collecting tweets for each entity as previously mentioned. However it is essential to note that we aggregate together tweets in a cluster for an entity that were published prior to the release date associated with the entity. All tweets post-release date referring to the entity are truncated. Further we ensure that the tweet cluster for each entity is rich in context, by enforcing a Jaccardian distance measure to compute a similarity threshold for candidate selection. A tweet may be a candidate if it has relevant information to contribute towards its associated entity. It is not restricted to being a direct reference to the entity but may be a reference to a parent company, a predecessor, an associated company etc. At the end of this process, we have clusters of tweets assigned for each individual entity in our knowledge base; if a cluster is an empty, we drop the entity from our knowledge base, due to lack of sufficient information to make a prediction.
Table 4.3: Snapshot of Tweet’s associated with iPhone 4S exemplifying the contextual information in associated entities.

<table>
<thead>
<tr>
<th>Date</th>
<th>#iPhone4S</th>
</tr>
</thead>
<tbody>
<tr>
<td>11 Dec 2011</td>
<td>Use Airplay to Watch Your iPhone 4S Camera on Apple TV [iPhone] <a href="http://goo.gl/fb/CtWrw,#tutorials">http://goo.gl/fb/CtWrw,#tutorials</a> #airplay #apple #camera #iphone #watch</td>
</tr>
<tr>
<td>11 Dec 2011</td>
<td>Getting the iphone on wed #teamiphone</td>
</tr>
<tr>
<td>11 Dec 2011</td>
<td>January 18th is marked on my calendar as a national holiday #iphone</td>
</tr>
<tr>
<td>11 Dec 2011</td>
<td>Switched On: A road trip with Siri: Each week Ross Rubin contributes Switched On, a column about consume... <a href="http://engt.co/uWKbDh,#iphone">http://engt.co/uWKbDh,#iphone</a></td>
</tr>
<tr>
<td>11 Dec 2011</td>
<td>The iPhone vs Android phones: What is your desired smartphone? Personally I think the iPhone is the better one a... <a href="http://bit.ly/tkTMO5">http://bit.ly/tkTMO5</a></td>
</tr>
</tbody>
</table>

The next stage is the feature engineering; as described we rely on three sets of vector features spaces and compare each of these features spaces to process the Twitter stream into a more manageable input to our SVR based machine learning algorithm. Once the feature engineering has been completed, we learn a model or rather three models in our case; which are capable of making predictions for these objects of interest extracted from future social media streams.

The last competent in the design of our system is the evaluation scheme. Like in most machine learning tasks we train our model on historical data/training set and then test on a subset that has been held out from the training cycle. More details on our evaluation scheme and the results from the Day-Zero review rating score estimator will be covered in the next chapter.
4.4 Summary

In this chapter we have described the mechanism behind our entire system. The pipeline begins at gathering data, building a knowledge base of facts - using Twitter streams to learn in a supervised setting. We talk in brief about the feature engineering driving our expectations of the system. We close with an overview of the skeleton of our system which relies on training a support vector regression model to estimate our Day-Zero review scores for objects closely associated with events identifiable in social media timelines.
Chapter 5

Results and Discussion

This chapter summarizes the bulk of the results derived from our work and goes on to describe the significance of these results. In addition we analyze some of the most interesting patterns and behavior exhibited in social media which offers some insight into the nature of our task.

5.1 Preliminary Tests

Previously, we talked about using support vector regression for the bulk of the learning task. The reasons that we chose to go ahead with this kernel method were both theoretical and empirical. In theory, since only a subset of the training set is used for learning, SVR offers immunity of a kind to over fitting and the problem of data sparsity. The independence of the number of features and the model is guaranteed - an overfitting protection that is better suited to handling large feature spaces. It also eliminates the need to perform any kind of feature selection - since SVM ensures that maximum of relevant features are set as support vectors.

Empirically other linear separators that we considered include firstly - linear regression, this regression is capable of learning the new relationships but limits performance to the number of variables; alternately the presence of outliers can have a huge impact on regression. Using variants such as ridge and lasso regression we were able to tone down the sensitivity of the system. We even considered the elastic net regularization scheme - a hybrid
combining both lasso and ridge regularization. The results of all the regression methods are summarized in the table.

<table>
<thead>
<tr>
<th>Regression</th>
<th>RMSE</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear Regression</td>
<td>0.1505</td>
<td>0.25001956</td>
</tr>
<tr>
<td>Lasso Regression</td>
<td>0.137</td>
<td>0.353</td>
</tr>
<tr>
<td>Ridge</td>
<td>0.1511</td>
<td>0.243659</td>
</tr>
<tr>
<td>Elastic Net</td>
<td>0.137</td>
<td>0.353</td>
</tr>
<tr>
<td>Passive Aggressive</td>
<td>0.209</td>
<td>1.064</td>
</tr>
<tr>
<td>SVR</td>
<td>0.128</td>
<td>0.180</td>
</tr>
</tbody>
</table>

Table 5.1: Results of initial training and testing on a probe sample data-set of 100,000 tweets. Different regression methods have been summarized with the best results from each.

As our probes on the sample subset of the data proved, performance of the SVR is comparatively more convincing and hence, it was the candidate chosen to perform learning for our task ahead. The results depicted here are in no ways indicative of any final conclusion - they simply strive to give an indication towards a fruitful approach on a developmental set.

5.2 Evaluation Scheme

In our learning framework, we refrained from segmenting into training and test sets since the data set was itself comparably small (as datasets for these tasks go). This is not necessarily a negative since minimizing the amount of training data for our model was an integral goal for our research. From a study of cross validation and bootstrapping methods (Kohavi, 1995), we decided to rely on a k-fold cross validation (rotational estimation) technique for estimating the accuracy of model. We iterated over 5-folds of our data set $D$, training each
time on $\mathbf{D/Dt}$ and tested on $\mathbf{Dt}$. The advantage of k-fold cross validation is of course that the variance dips with stratification - k-fold essentially averages variance over the specified k-folds, minimizing performance to sensitivity of partitioning.

To estimate accuracy/concurrently the error for the system, we have essentially used the RMSE or root mean squared error metric. The MSE is defined as the mean difference in squared error, summed over all samples or mathematically as:

$$MSE(t) = \frac{1}{N} \sum_{i=1}^{N} (y(x_i) - t_i)^2$$

here T is a sample set of test data with in sample data points. For each $i$ in T, $y(x_i)$ gives predicted review score and $t_i$ the true score from our knowledge base. In alternative schemes a certain window for error is also allowed, however since our framework predicts the base review score for an object (i.e. 9 in lieu of 9.5) we have abstained from any degree of allowable error in evaluation. Consequently, the RMSE is reported as the square root of the mean squared error.

Alternatively, we have also reported the coefficient of determination - another statistical measure which reports the degree of fit of model. Close fit or a 'good fit' can be treated to indicate the tendency of the system to over fit. The reason this is undesirable off-course is that for a model that is dependent on a live stream of data it is perhaps illogical to make any assumption on the nature of the incoming data. If our model was fit well to the current trends and features on Twitter, it would tend to be less and less accurate with changing times and trends. Evaluating coefficient of determination ($R^2$) gives an idea of degree of fit in addition to qualitative performance of the model (described by its accuracy). The closer a value is to 1, greater is the goodness of fit. Mathematically the $R^2$ measure is defined as:

$$R^2 = 1 - \frac{SS_{rest}}{SS_{tot}}$$

where, SS rests is the residual sum of squares defined as
\[ SS_{\text{rest}} = \Sigma(y_i - t_i)^2 \]
denoting sum of squared errors and \( SS_{\text{tot}} \) is traditional variance defined as
\[ SS_{\text{tot}} = \Sigma(t_i - t)^2 \]

5.3 Outcome

5.3.1 Results

The outcome of the results from following the 5-fold cross validation of our system are summarized in the Table 5.2 below. Among the automated regressors evaluated: the baseline performance is given by a simple statistical approach of predicting the median score for each entity. The median is decided over the entirety of our knowledge base individually for each entity category. As depicted in the table; all of the evaluated automated predictors outperform the baseline by a significant margin.

The product data-set is of particular interest as the task of extracting consumer products from social media streams has perhaps been undertaken rarely in the past. The results for product streams follow quite intuitively; the results from the bag-of-words feature space have been improved upon by the TF-IDF weighting and the \( POS_{\text{nouns+verbs}} \) restricted space. This holds with our hypothesis, that for products at least; nouns and verbs are highly descriptive of individual product-related tweet clusters and incorporate a lot of data in a concise feature space. It is also noteworthy here, that considerable care was taken by the author’s in curating the product knowledge base, over the others. Following, the ease with which product mentions are identified and extracted from social media, without ambiguity suggests that the \( POS \) feature space is the promising direction for further analysis and representation of products and relative information from streams.
<table>
<thead>
<tr>
<th></th>
<th>Support Vector Regression</th>
<th>Elastic Net</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RMSE</td>
<td>$R^2$</td>
</tr>
<tr>
<td><strong>Products</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline</td>
<td>0.101267</td>
<td>0.009847</td>
</tr>
<tr>
<td>Bag-of-words</td>
<td>0.086502</td>
<td>0.712217</td>
</tr>
<tr>
<td>TF-IDF</td>
<td>0.084401</td>
<td>0.538556</td>
</tr>
<tr>
<td>$POS_{verbs+nouns}$</td>
<td><strong>0.083502</strong></td>
<td>0.596987</td>
</tr>
<tr>
<td><strong>Movies</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline</td>
<td>0.109473</td>
<td>0.0175973</td>
</tr>
<tr>
<td>Bag-of-words</td>
<td><strong>0.092431</strong></td>
<td>0.142821</td>
</tr>
<tr>
<td>TF-IDF</td>
<td>0.0955809</td>
<td>0.227881</td>
</tr>
<tr>
<td>$POS_{verbs+nouns}$</td>
<td>0.0946654</td>
<td>0.198728</td>
</tr>
<tr>
<td><strong>Games</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline</td>
<td>0.1705435</td>
<td>0.141258</td>
</tr>
<tr>
<td>Bag-of-words</td>
<td><strong>0.144680</strong></td>
<td>0.007938</td>
</tr>
<tr>
<td>TF-IDF</td>
<td>0.146364</td>
<td>0.015284</td>
</tr>
<tr>
<td>$POS_{verbs+nouns}$</td>
<td>0.147011</td>
<td>0.0242890</td>
</tr>
</tbody>
</table>

Table 5.2: RMSE and coeff. of determination for each model in our research: automated predictors evaluated against 5-fold cross validation with results reported for highest accuracy across all folds. Different feature models and learning algorithms have been compared.

For the movie task we observe that the performance of the bag-of-words regressors exceeds performance by either the $POS$ approach or the TF-IDF weighting scheme in the case of SVR. While this may seem surprising, a possible explanation is the difficulty of disambiguation in tweets regarding movies; in our approach we have refrained from including tweets that may be related to a review entity such as chatter around an associated actor/director/production which could provide context an information around the topic.
This may lean towards indicating that feature spaces such as POS and TF-IDF for this task; respond much better when context around an entity is provided. The Elastic Net regressor; performs better with the additional feature spaces, with performance boosted to a great degree by POS tags in this learning mechanism. This result is indeed surprising and we attribute this huge margin gain to the ability of Elastic Net to perform simultaneous regularization and variable selection. With matrices for condensed features in this review class being highly sparse we anticipate Elastic Net regression methods to perform better on the Movie set than Support Vector Regression.

<table>
<thead>
<tr>
<th>Review Class</th>
<th>Extracted Entities</th>
<th>Median Scores</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consumer Products</td>
<td>49</td>
<td>7.7</td>
<td>1.007</td>
</tr>
<tr>
<td>Movies</td>
<td>207</td>
<td>6.3</td>
<td>1.085</td>
</tr>
<tr>
<td>Games</td>
<td>235</td>
<td>8.0</td>
<td>1.596</td>
</tr>
</tbody>
</table>

Table 5.3: *Table describing useful statistics for each class in our knowledge-base.*

The Video Game data-set is pretty similar to the movie data-set in the aspects that

- it contains, large number of entity mentions extracted from 15 million tweets

- again, most of these entities have sparse features

What makes this quite different however is that despite being the largest set it has the highest median review score for entities and greatest standard variation. Knowing this, it is quite easy to see why RMSE is higher for the games class over all the others. What’s interesting to observe is that we see similar scores between different feature models and learning approaches. Like results for the movie data-base it achieves best SVR performance on the bag-of-words model and best Elastic Net performance on the POS model. Again, with the similarity to sparse features like the movie database; it is reinforced that there must
definitely be a co-relation between cluster size/ feature sparsity and prediction accuracy for each review entity. In the case of games again; it is difficult to disambiguate references to actual game title; since a lot of game titles are common world named-entities which could refer to movies, comics and books. With the added difficulties of extraction in this domain we are encouraged by the uniformity of performance on the task of learning to predict game review scores.

5.3.2 Observable Patterns

While considering features for our models one of the points of interests was to explore effects of introducing cross channel signals to the lexical features in our results. (Bollen et al., 2011) for instance indicates the pattern between stock indicator movement and general sentiment on Twitter. Sentiment has a varied effect in different tasks, for our purpose however we chose to ignore its potential benefits in light of the drawbacks of such a system. Reasons for these are numerous - firstly the granularity of sentiment is at Tweet level; while incorporating Tweets over clusters they normally even out to give a balanced overall sentiment which in sum, provides little to none relevant information. Secondly, in relation the information encoded into our model includes related yet indirect mentions - in light of stronger context analysis we did not want to consider the implications of off-topic Tweet mentions offsetting polarity for the entire cluster. Neither were we keen on discarding off-topic Tweets which potentially contributed consequential signals to the object in consideration. Third, sentiment classification being a challenging problem in the Twitter domain (an unsolved problem), introducing Tweet polarity would just lead to propagation of error in the system. Finally we established our goal in defining this framework to be independent of cross channel information as it was our (the authors) intention to design a system that could be truly pluggable to the stream of data and make predictions for Day-Zero review scores without being genre specific. There is potential for cross channel information that is generic,
however in most situation requires dependence on genre domain specific agents. As we have elucidated in building our knowledge base-the task of specialized genre-based extractors is a challenging task. We would like to avoid this hurdle in deploying such systems by limiting their dependency purely on information extractable from live social media feeds.

For each of our review entity categories we have summarized the value of the regression coefficients in Table 5.4.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Coefficient</th>
<th>Feature</th>
<th>Coefficient</th>
<th>Feature</th>
<th>Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>vox</td>
<td>0.01127</td>
<td>recapped</td>
<td>0.02473</td>
<td>assassinate</td>
<td>0.02208</td>
</tr>
<tr>
<td>kobo</td>
<td>0.00449</td>
<td>wind</td>
<td>0.02448</td>
<td>japan</td>
<td>0.01911</td>
</tr>
<tr>
<td>200</td>
<td>0.00356</td>
<td>breaking</td>
<td>0.02383</td>
<td>konami</td>
<td>0.01910</td>
</tr>
<tr>
<td>tablet</td>
<td>0.00353</td>
<td>risk</td>
<td>0.02356</td>
<td>enemies</td>
<td>0.01903</td>
</tr>
<tr>
<td>launched</td>
<td>0.00340</td>
<td>energy</td>
<td>0.02097</td>
<td>ric</td>
<td>0.01817</td>
</tr>
<tr>
<td>priced</td>
<td>0.00273</td>
<td>harnessing</td>
<td>0.01952</td>
<td>zen</td>
<td>0.01756</td>
</tr>
<tr>
<td>new</td>
<td>0.00247</td>
<td>dreamed</td>
<td>0.017889</td>
<td>vita</td>
<td>0.01710</td>
</tr>
<tr>
<td>color</td>
<td>0.00231</td>
<td>storage</td>
<td>0.01441</td>
<td>pac</td>
<td>0.01691</td>
</tr>
<tr>
<td>sell</td>
<td>0.00219</td>
<td>duke</td>
<td>0.01305</td>
<td>revenge</td>
<td>0.01645</td>
</tr>
<tr>
<td>nook</td>
<td>0.00210</td>
<td>gas</td>
<td>0.01302</td>
<td>wes</td>
<td>0.01618</td>
</tr>
</tbody>
</table>

Table 5.4: Highest regression coefficient for each review entity class along with corresponding feature.

It is interesting to observe that the top coefficients for the product and games classes follow a pattern while the movies category is somewhat more random. In the former two classes - the features with highest regression coefficients follow trend of being descriptive factors of the entity - named-entities associated with the entity itself, a parent entity such as a product_manufacturer or game_studio. In addition, the trend also shows an inclina-
tion towards verbs and adjectives being loosely included into this set of features that we loosely term as descriptive factors. This bears further weight on our initial conclusion that $POS_{verbs+nouns}$ in themselves contribute important signals towards regression of product scores. In addition we observe that product modifiers such as physical attributes and numerical values also contribute significantly towards the model.

The random form of high regression coefficients in movies is most likely attributed to dimensionality of our feature space in this problem. The observations drawn from sampling the highest coefficients leads us to believe that the context around movies in tweets is perhaps limited and may not be provide strong information towards our predictive task. Tweets around movies often suffer from ‘sensationalization’ which may result in mis-information being added around the context. Another common factor is the ‘spoiler-effect’: where writers and micro-bloggers may minimize the textual information in a tweet - in order to avoid giving away details relevant to the plot of the movie. Reviews of products and video games rarely suffer from this phenomenon as the effects of revealing spoilers is minimal in the case of their domain.

We are able to hence predict review scores across these genres and in the process extract ‘critical’ measures around the framework that solidify belief in the system. The slight disparity in the way features contribute across genres indicates that the approach is generic enough that we may achieve high accuracy in predicting base scores but for a finer-grained task - our framework is perhaps lacking in genres that enjoy a larger attention on social media increasing the dimensionality of the problem space. For such domains i.e. movies we would suggest the use of additional surface features extracted from other channels used in aggregation to enhance results on the task; however that should be discouraged for a more generic approach.
Chapter 6

Conclusion

Over the course of this thesis, we made an effort at building a framework capable of extracting available online facts about products, movies and games; given a database of the objects. With a knowledge-base generated by scraping the Internet, we are then able to extend the framework to learn to predict prospective review scores for future objects pertaining to the above-mentioned classes; we predict review scores prior to the release of the product itself-terming these review ratings as *Day-Zero* rating scores. Our contribution to this task is the ability to predict the scores a sufficient period prior to the release date; given sufficient amount of information available from Twitter (social media) streams. This answers an interesting question in regards to the task of making prospective predictions with data from social media feeds. Our results definitely prove that with existing learning frameworks and natural language processing systems we are able to achieve sufficient degree of accuracy for the task. Furthermore these predictions are independent of any cross channel information and depend purely on the stream of data being fed-in. We sufficiently establish the advantage of such a system and show that it performs somewhat consistently across genres. However, the system does also provide certain insights into the practicality of such tasks. We achieve a satisfactory level of accuracy predicting review scores in the domain of consumer electronics product and video games; however the task is more drawn out and vulnerable to the effect of high-dimensional data in the case of predicting movie review scores. Regardless we are able to achieve high levels of accuracy in predicting a base score for entities across all three domains.
We’ve observed trends that are interesting and offer prospective avenues of study for future work. One of the possibility for extension is to experiment with cluster sizes for each object and study its effect. A promising train of thought is in regards to timeseries analyses which has proven results in tasks of linking Twitter sentiment to opinions polls (O’Connor et al., 2011). Coupled with timeseries analyses - we would very much like to identify patterns in time windows of maximum sensitivity. (Asur and Huberman, 2010) talk about critical periods prior to release of the movie when campaigns are in full swing - this being considered the time when majority of the opinions are propagating around Twitter chatter. It is of quite some interest to the authors - in exploring the possibility of defining an analogous critical window to analyzing volume of chatter around reviewable entity and categorizing uptrends and downtrends. If this information could be encoded into our model - it would have been of worth to discover the existence of a correlation.

Another interesting train of thought is the side-effect of studying difference in linguistic trends between Day-Zero Reviews and post Day-Zero reviews i.e. expert reviews published on dedicated channels. As discussed above the ‘spoiler effect’ and other phenomenon in short messages ensure that there is a certain degree of time for which Tweets may still be illusive in domains such as movies. For products and video games it is quite the opposite where - after release there is a storm of tweets that are highly descriptive of the entity in question. It would be interesting to see what are the trends in linguistic patterns in both Tweets and actual review text at this stage and how much they vary from linguistic information presented through our system. Apart from linguistic cues, there are also other indicators that beg to be examined - how does volume of tweets per entity look like for instance?

Finally, with growing sensitivity of public over mining of personal data - the practice of data mining has seen its fair share of controversy. One of our motivations towards a single-
channel information model was to be able to learn without violating privacy concerns. Our framework is inclined towards knowledge mining rather than capturing any information about the user itself. We refrain from any form of network analyses or graph-based relationship modeling despite the public nature of this information. The framework is capable of extracting and storing knowledge-volume of chatter over topics, extractable events, linguistic (lexical analysis) and polarity if necessary. It is with the underlying philosophy that as data mining evolves to a more user conscious knowledge mining; systems such as the one proposed here will become the norm in the social web mining approaches and downstream applications.
References


[23] Rodak, J., Xiao, M., & Longoria, S. Predicting Helpfulness Ratings of Amazon Product.


