“Electronic Devices Sales Prediction Using Social Media Sentiment Analysis”

Sahar Nassirpour, Parnian Zargham, Reza Nasiri Mahalati

Introduction:

Nowadays, social media has become a platform for people to convey their voice to the public. Among various opinions that people share and exchange, there are a lot of comments about consumer products. Recently, it has been shown that the chatter of the consumers in the social media such as Facebook, Twitter, MySpace, Google+ and etc. correlates strongly with the product’s actual financial performance in the market. This forms a beneficial database for companies to analyze the consumers’ demands in order to make a quantitative prediction of their potential customers. We were inspired by this new potential and decided to exploit our machine learning skills in order to investigate the trends in the social media discussions and use them to predict the actual sales of electronic devices before their release.

We chose the electronic device market, because first of all, the electronic device market has always been very competitive, and recently with the growing variety of new products it is very challenging to predict which brand will dominate the market. Having an estimate of the product sale before the release of the product would provide the company with prior knowledge of its profit and also help them decide the quantity of the release in different regions based on the request. Secondly, the actual outcome of the electronic device market sales can be easily accessed and we can use it to train our model.

In this project, we try to predict the sales of electronic devices based on the sentiment of the comments made about them on Twitter before their release. To predict the sentiment of the comments we use a machine learning framework based on recursive autoencoders (RAE) for sentence-level prediction of sentiment label distributions. This method learns vector space representations for multi-word phrases. In sentiment prediction tasks these representations outperform other state-of-the-art approaches on commonly used datasets, such as movie reviews, without using any pre-defined sentiment lexica or polarity shifting rules. After sentiment prediction, we use linear regression with four features to predict the sales of the devices. We have selected *total number of comments, number of positive comments, total number of retweeted comments* and *number of retweeted positive comments* as the four features of our model. We compare our prediction accuracy with a linear regression only based on the total number of comments with no sentiment analysis and show that our method significantly improves the prediction accuracy.

In the following sections, we first describe what kind of data we plan to use for our purpose and how we have collected them. Next, we will talk about the techniques that have been used to achieve the goal of predicting the sales. And finally, we talk about the results of our projects and come to a conclusion.

Data:

For this project, we need two sets of data: first the social media content that will be used to form the input features to our system, and second the sales revenue numbers of different electronic device products that will be used as the target variables. Together these two datasets will form our training set.

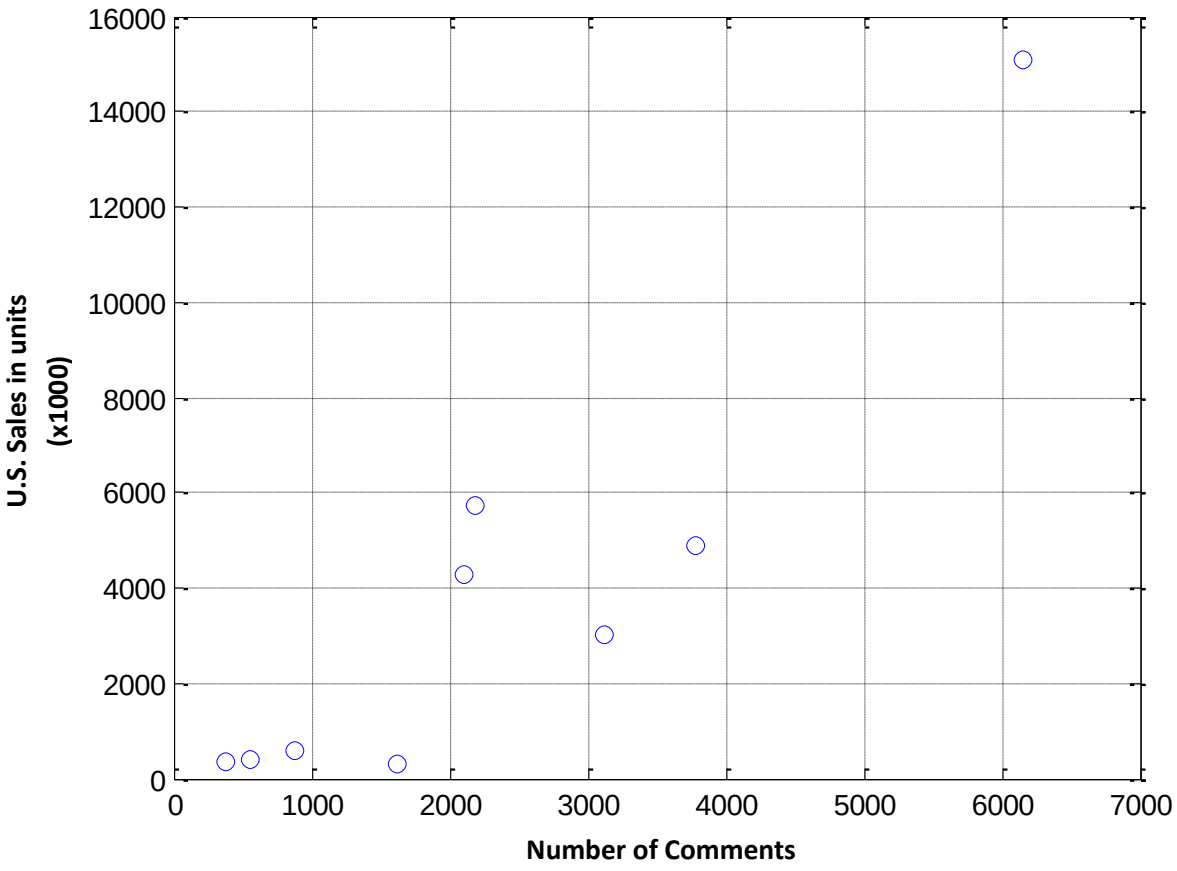
- ***Social Media content:*** For the purposes of this project, we decided to use Twitter data. We have collected Twitter data through the http://stanford.discovertext.com/ website. The way the website works, is that you can provide it with a specific keyword and it automatically fetches all the tweets that contain that keyword. It also allows for time windowing which is very helpful for our project and enables us to focus only on the comments that are posted in the pre-release time window of a specific product. Next, we need to modify the data to have a consistent format. In particular we need to parse the data and convert it to a readable format for our algorithm.

  The format that was needed for our project dictated us to map the comments into a vector space. Basically, we stored each comment in a binarized vector, each component being the number of that word in the comment in a predefined body of words. A MATLAB function was written that reads all the comments and uses MATLAB's wordMap functions to find that word in our pre-stored body of words and store the vector representation of all the comments.

- ***Product Sales data:*** We picked 9 electronic devices that have been recently released and collected the actual sales numbers from the internet [4].

The following table summarizes all the collected data. And the following figure shows the data in a plot.

Product Make and Model	First Quarter after Release	Sales in units (figures in thousands)	Number of Twitter Comments
Apple iPhone4	3rd quarter 2010	4912	3777
Apple iPhone4S	4th quarter 2011	15073	6155
Apple iPad2	2nd quarter 2011	4293	2106
Apple iPad3	2nd quarter 2012	5713	2189
Samsung Droid Charge	3rd quarter 2010	300	1629
Samsung Galaxy S	1st quarter 2011	354	371
Samsung Galaxy SII	3rd quarter 2011	574	882
Samsung Galaxy SIII	3rd quarter 2012	3000	3122
Samsung Nexus	2nd quarter 2011	401	551



Methods:

For this project, we decided to use the "Semi-Supervised Recursive Autoencoders for Predicting Sentiment Distributions" [1] method. The reason being that the mentioned algorithm has several desirable properties. First of all, it takes the whole sentence including the syntax and grammar into account instead of just focusing on the bag-of-words. This is done through paying attention to the order of words in the statement, which gives us a more accurate evaluation of the sentiment behind the statement. Second, it does not require any predefined dictionary of positive or negative words.

After performing the sentiment analysis, we extract four features from all the gathered Twitter data for each product and try to fit a model to our data that best predicts the sales figures of the electronic devices.
The following sections describe each of these steps:

Sentiment analysis
As described above, after gathering the data, we first used the "semi-supervised recursive autoencoders for predicting sentiment distributions" method developed by Socher et al to classify the comments about each product into positive and negative categories. For training purposes, we manually classified 2000 tweets about iPhone5 into positive and negative groups and used them as training data in our classifier (iPhone5 was chosen because it had the most number of comments both positive and negative and so was ideal for making ground truth).

Using 70/30 cross-validation on this data, we found out the comment classification accuracy to be 83%. We then used the trained classifier to categorize the comments about other products into positive and negative groups.

Sales prediction
After sentiment analysis, we used a linear regression model with four features to predict the sales of the products in table 1. We used the following four features suggested by Asur and Huberman [2] in our model: total number of comments, number of positive comments, total number of retweeted comments and number of positive retweeted comments.

We used the sales data for iPad2, iPhone4, Samsung Droid Charge, Samsung Galaxy Nexus, Samsung Galaxy S, Samsung Galaxy SII and Samsung Galaxy SIII to train our model and cross-validated it on iPad3 and iPhone4s sales.

Experiments and Results:

We tested our sales prediction accuracy against a linear regression model that only uses the total number of tweets about a product with no sentiment analysis and added features. The table below shows the results. It is seen that our method significantly improves the prediction accuracy.

Product	Pred. error using linear regression and no sentiment analysis	Pred. error using our model
iPad3	56%	35%
iPhone4s	48%	30%

Conclusion:

In this project, we showed how the content of social media produced by the mass public can be used to predict how well a product is going to do in the market before its release. In particular, using the Twitter data about certain electronic devices, we constructed a model that can forecast the actual sales figures of a product with a reasonable accuracy. We also performed an state-of-the-art sentiment analysis on the comments to determine the polarity of those comments without using a predefined lexica of labeled words.

References:

1. R. Socher et al., "Semi-Supervised Recursive Autoencoders for Predicting Sentiment Distributions",EMNLP '11 Proceedings of the Conference on Empirical Methods in Natural Language Processing.

2. S. Asur et al., "Predicting the Future With Social Media",arXiv:1003.5699.

3. R. Sharda et al., "Forecasting Box-Office Receipts of Motion Pictures Using Neural Networks", CiteSeerX 2002.

4. http://www.businessinsider.com/apple-and-samsung-just-revealed-their-exact-us-sales-figures-for-the-first-ever-time-2012-8