CS4022 Social Media Mining

Individual Project 2:

Amazon Software Review Mining

By: Tan Ching Fhen **Mat ID: U1920787D**

1 Introduction:

The Amazon Review Data [1] is a rich source of review data for companies to perform social media mining. Utilizing social media mining, we can extract many insightful information that can be used to enhance business processes. In this project, I have decided to perform social media mining on *software* reviews because many of the largest companies in the world like Microsoft are driven by software.

Using many social media mining techniques such as topic modelling (Section 3.1), word frequency mining (Section 3.2), sentiment categorization (Section 3.5), review semantics (Section 3.3), keyword extraction (Section 3.3), and information retrieval (Section 3.6), I hope to extract insightful information from *software* reviews, explore better techniques, and uncover potential areas for social media mining in future.

All plots and figures are illustrated in the appendix.

2.1 Data Description:

In the Amazon Review Data, under the *small* datasets section, I downloaded the *software* reviews section as well the meta-data. In each sample, the fields that I utilized are “overall”, “summary”, “reviewText”, “asin” and “brand”. The “overall” field refers to the rating given to the product which ranges from 1 to 5, “asin” field is an identifier, “summary” field is the summary of the review and “reviewText” is the full review. I decided to include the “summary” field because it is much more concise, which I hypothesize that it will lead to better features for sentiment categorization. In Figure 17 and 18, I show a sample of the raw data.

2.2 Preprocessing:

With that being said, I will now describe the basic preprocessing steps that I performed to make the data suitable for my social media mining techniques.

First, I merged the review (Figure 17) and metadata (Figure 18) by the “asin” field. I assigned, reviews with ratings 1-2 a negative sentiment, reviews with rating 3 a neutral sentiment and reviews with ratings 4-5 a positive sentiment. Some of these samples are displayed in Figure 16. Furthermore, I balanced the response variable, “sentiment” (Figure 19). Introducing a neutral sentiment and balancing the response classes will make the sentiment categorization task more difficult but retains a lot of information that can tell us, “Why a customer gave neutral feedback?”, and “How can we improve our business processes so that customer satisfaction can improve from neutral to positive?”.

Next, I performed all basic processing steps such as duplicate removal, lower-casing, stop word removal, punctuation removal, lemmatization (*textstem* R package [10]) and encoding of the response variable by the mapping {“NEU”:0, “POS”:1, “NEG”: -1}. Note that depending on the type of mining technique used, the processing step may be different to suit that task. For instance, I performed lemmatization, stop word removal and punctuation removal for *TFIDF* mining (Section 3.2), but not for embedding generation (Section 3.3) because sentence semantics need to be retained for *Transformer* models – these language models inherently understand punctuations, grammar, and meanings of words.

In total, the processed dataset has 4440 samples.

3 Social Media Mining Techniques:

3.1 Topic Modelling:

Topic modelling is an unsupervised task of assigning hidden topics to textual documents. There are many topic models available, each with their strengths and weaknesses. Perhaps the most state-of-the-art and handy of those is *Top2vec* (R package [2]), which has been found to produce topic words that are much more informative (Figure 25) as found by Dimo Angelov [3]. In addition, in contrast to Latent Dirichlet Allocation (LDA), *top2vec* does not require the number of topics to be specified beforehand. Thus, I will utilize *top2vec* [2] over LDA.

Top2vec uses Doc2vec to produce word and review embeddings, then UMAP (a dimensionality reduction technique) is performed, followed by clustering using HDBSCAN. These clusters produced are the hidden topics and the topic words are the word embeddings that are closest to each topic centroid. With that said, in addition to assigning each review the most similar topic, I also computed the dot product similarity to each topic using the embeddings (Figure 26) – this is similar with the topic distributions generated by LDA. This produces much more informative features for sentiment categorization models later (Section 3.5) – having a degree of association to topics is more informative than simple yes or no association.

There were 36 topics found in total, and Figures 8-13 show the topic words for 6 of these topics – the larger the word, the more relevant the word is to that topic. Figure 8-10 are the most important topics while Figure 11-13 are the least important topics (I will describe topic importance later in Section 3.5). The most important topic, topic 35, have topic words like “excellent”. Looking at the distribution of sentiments by topic (Figure 24), we can observe that topic 35 has mostly, positive sentiments. On the other hand, topic 6 (Figure 9), has topic words like “junk” and it has mostly negative sentiments (Figure 24). This means that, the more unbalanced the sentiment distributions for the topic, the more informative the topic (in predicting sentiment) and the corresponding topic words.

With this knowledge, we can also find out, for a particular review, product, or brand, what is its most similar topic (by finding the topic with the highest similarity score in Figure 26)? What is the sentiment distribution towards that topic (Figure 24)? What (topic) words are usually used to describe that review, product, or brand (Figures 8-10)? For instance, for sample 49 (Figure 22) about a Microsoft product, its most similar topic is topic 6, which has many negative topic words (Figure 9). For sample 828 (Figure 23) about an Apple product, its most similar topic is topic 35, which has positive sentiments and associated words (Figure 8).

3.2 TFIDF Mining:

Term Frequency and Inverse Document Frequency (TFIDF) is a numerical statistic that reflects how important a word is in the review corpus. I computed TFIDF features using the text2vec R package [4]. However, TFIDF features are very sparse, and high in dimensionality. This makes learning more difficult especially for certain models due to the curse of dimensionality. Thus, I try to reduce this effect by taking only the most frequent topic words found in Section 3.1, which tend to be more informative and higher in variance (this reduced the total number of TFIDF features by over 10000).

Nonetheless, we can still observe that TFIDF are poor predictors of sentiment (Figure 2) with very low feature importance (I will describe feature importance in Section 3.5). Figure 3 shows some of the most important TFIDF features like “easy” and “great”, while Figure 4 shows some of the least important TFIDF features.

Since TFIDF features are generally poorer predictors (Figure 2), I only used unigrams, because more n-grams likely would not make significant contributions.

3.3 Review Embeddings:

Semantic embeddings are vectors that represent meanings. Specifically, I will use contextualized embeddings as additional features for my sentiment categorization model. Contextualized embeddings are dynamic; depending on context, the same words can be assigned different embeddings. This accounts for situations where the same words have multiple meanings.

To generate contextualized embeddings, I made use of *Transformers* (R package [5]), that are state-of-the-art language models that utilize self-attention mechanism. Specifically, I generated embeddings for the "summary" field and "reviewText" field separately. The reason being that the summaries are generally more concise and less noisy, thus I hypothesize that they would produce better embeddings for sentiment categorization in Section 3.5.

Because the "reviewText" field tend to be very long and noisy, I used Rapid Automatic Keyword Extraction (RAKE) [6] to extract important keywords. These keywords are passed through the *Transformer* models to generate another set of keyword embeddings. The motivation here is to reduce the impact of noise and hopefully improve embedding quality.

In figure 3, we can observe that the summary embeddings are most important in predicting sentiment i.e., informative, while the review and keyword embeddings are less so (possibly because the keywords generated by RAKE are still quite noisy, with a lot of redundant keywords). This validates my hypothesis that, a summary of a review is a better source of raw data compared to a full review.

3.4 Additional Feature Engineering:

In this section, I will describe some additional features that I generated manually, which may help improve sentiment categorization.

First, I extract the number of words and number of uppercase words (I group these features as "other_features" in Figure 2). The motivation is that the number of uppercase words or words may have correlation to the sentiment of the review.

Next, I extract the number of punctuations. Punctuations such as exclamation marks may correlate with customer sentiment, thus it's useful to retain such information. In figure 5, we can observe the feature importance for punctuation; commas, full stops and exclamation appear to be somewhat informative.

Lastly, I also utilized sentiment lexicons. Specifically, I use WKWSCI Sentiment Lexicon Version 1.1 [7]. For every review, I acquired the sentiment scores for all the words that were also in the sentiment lexicon, then computed the summation and normalized by the length of the review. This means that if the total sentiment lexicon score is positive, there are more positive sentiment words than negative sentiment words in the review. In addition, for reasons stated before, I computed the total sentiment lexicon scores for "summary" and "reviewText" field separately. Intriguingly, sentiment lexicon scores were the best features in predicting sentiment category, far more so than all other features (Figure 1 and 2).

From Figure 2, we can observe that both these features are effective in predicting sentiment. This shows that if we have some domain knowledge or some kind pre-constructed information (such as sentiment lexicons), it is effective to leverage them.

3.5 Sentiment Categorization:

This chapter will describe the models trained for sentiment categorization and how I acquired the feature importance scores.

There are three sentiment categories, POS, NEG, and NEU, which are equally distributed (Figure 19) as described in Section 2.2. I split this dataset into training and test sets by the ratio 7:3.

I trained 5 types of models namely: lightgbm [8], random forests, extra trees, naïve bayes, and support vector machines (SVM trained for multi-class classification using one-to-one scheme) [9]. For all models, I used the default hyperparameters for fair comparison.

In total there were 2223 input features, consisting of "sentiment_lexicon" features, "other_features", "topic_similarity" features, embedding features, punctuation features and TFIDF features (Figure 2). However, due to training time concerns, I only used the top 300 features with the highest importance scores. These feature importance scores (Figures 1-7, 14) were derived by training lightgbm [8] – the higher the feature importance, the better it is in predicting sentiment.

For SVM, I also trained a reduced model, "SVM-reduced", that only uses the best features, "sentiment_lexicon_score" features. I hypothesized that the large number of features might have created a lot of noise which tends to cause issues for SVM learning.

Figure 15 shows the accuracy, precision, recall and f1 scores. We can observe that lightgbm, performed the best. Thus, I performed further hyperparameter tuning using grid search (lightgbm-tuned). Also, we can observe that "SVM-reduced" performs better than SVM even though it uses only 2 features (compared to 300).

From the confusion matrix and statistics (Figure 27 and 28), we can observe that neutral sentiment (NEU) is more difficult to predict – the sensitivity and accuracy is much lower for this class. To give an example, figure 29 shows a sample that is neutral, but the best model predicts a negative sentiment – which is understandable because the word "leery" has a negative connotation. On the other hand, figure 30 shows a sample that is neutral, but the best model predicts a positive sentiment – which is also understandable because of the words "good" and "easy".

3.6 Information Retrieval

Lastly, this chapter will describe how I utilized the *top2vec* model and embeddings (Section 3.1) to perform some interesting semantic retrieval tasks. For instance, given a *new* review, what is the most similar topic? Given some keywords, what is the most similar review? When the number of reviews and topics get very large, these tasks will have some utility, thus is worth exploring.

Given a *new* review that was not found in the dataset in Section 2.2, we can efficiently assign it to an existing topic without having to retrain the topic model by doing a semantic retrieval. In detail, the *top2vec* model first assigns the unseen review, an embedding **r** of size 1 x 300. I performed matrix multiplication with the existing topic embedding matrix $\mathbf{T}^{\mathbf{T}}$ of size 300 x 36 to produce a vector of similarity scores. The most similar topic for the review shown below was topic 33 (Figure 32), which happens to be about video editing:

I am a total amateur when it comes to editing but needed to try something for editing the photos I take for my business. This software was a quick and easy download. Once downloaded, I was able to start using it pretty quickly, the learning curve was one that I could understand without too much trial and error. I take many photos of different items in the same location, so when I need to edit, it's generally the same thing which needs editing, this software is perfect for that. N ow I can fix multiple photos at one time, saving a lot of time! The price is great if you're not sure about the program, not a huge expense like many others out there. For me, and my level of ex pertise, or lack of it, this program fits the bill!

Next, given a short query, "useful tool to compress file", I can retrieve a similar *review* from the dataset using the same steps. First, compute the query embedding **q** using the *top2vec* model and perform matrix multiplication with the review matrix $\mathbf{R}^T$ of size 300 x 4440. This produces a vector of scores, and the most similar review happens to be one about win-zip (Figure 31).

4 Discussion:

This section will describe some further learning points, and insights found that can be used to improve business processes and mining tasks in future.

From figure 2, we can observe that TFIDF features are least useful in predicting sentiment likely because of the sparsity and high dimensionality. The number words and uppercased words ("other_features"), performed surprisingly well. This shows that, applying our domain knowledge to manually extract useful features can be quite effective. "Topic_similarity" features (Figure 26) also performed quite well. Sentiment lexicon features are surprisingly effective as features for sentiment categorization. With that said, future mining tasks can focus on the best features found and reduce development time.

Furthermore, based on my findings (Figure 1 and 2), we can observe that the summaries of reviews reveal the most information of customer sentiment and satisfaction. Therefore, when collecting customer reviews in future, it will be effective to ask for a short summary of the review if not already done so.

In addition, based on my experiments in Section 3.1 and 3.6, we can observe the effectiveness of topic modelling using *top2vec*. Using the embeddings generated, we can perform interesting tasks like semantic information retrieval like I've shown. In addition, we can extract many informative topic words as I've shown in Figures 8-13.

Finally, I have shown that we can use topic modelling in conjunction with sentiment categorization to improve business goals. Let's say we want to find out how we can improve customer satisfaction from neutral to positive or negative to neutral, we could do the following: First, observe the sentiment distribution of topics (Figure 24), we find that topic 6 has a high proportion of negative sentiment while topic 23 has a high distribution of neutral sentiment. Next, bring out the topic words from those topic (Examples in Figures 8-13) – this will summarize the key areas to improve our software product. Finally, bring these insights to the relevant teams, where they can focus on solving those problems that customers are facing.

5 Conclusion:

To conclude, I've made use of many social media mining techniques namely: topic modelling, sentiment categorization, TFIDF, sentiment lexicons, contextualized embeddings, information retrieval, and feature engineering.

In future work, we can perform more in-depth analysis of the topics and topic words. We could develop a dashboard that summarizes the topic words associated to each brand, making analysis easier. Furthermore, we could automatically or manually assign sentiments to the topic words to create a *software domain specific sentiment lexicon*; for topics like 6 and 35 that have very unbalanced sentiment distribution (Figure 24), we can reliably assign the most relevant topic words like "junk", "crash" and "reinstall" (Figure 9) with high negative sentiment score, while "excellent", "paint-shop" and "exposure" can be assigned high positive sentiment score. Since we learnt earlier that sentiment lexicons make very good features, this can have some utility.

6 Appendix:

Here, I append all plots and figures generated from R packages such as *ggplot*, *wordcloud2*, and *caret*.

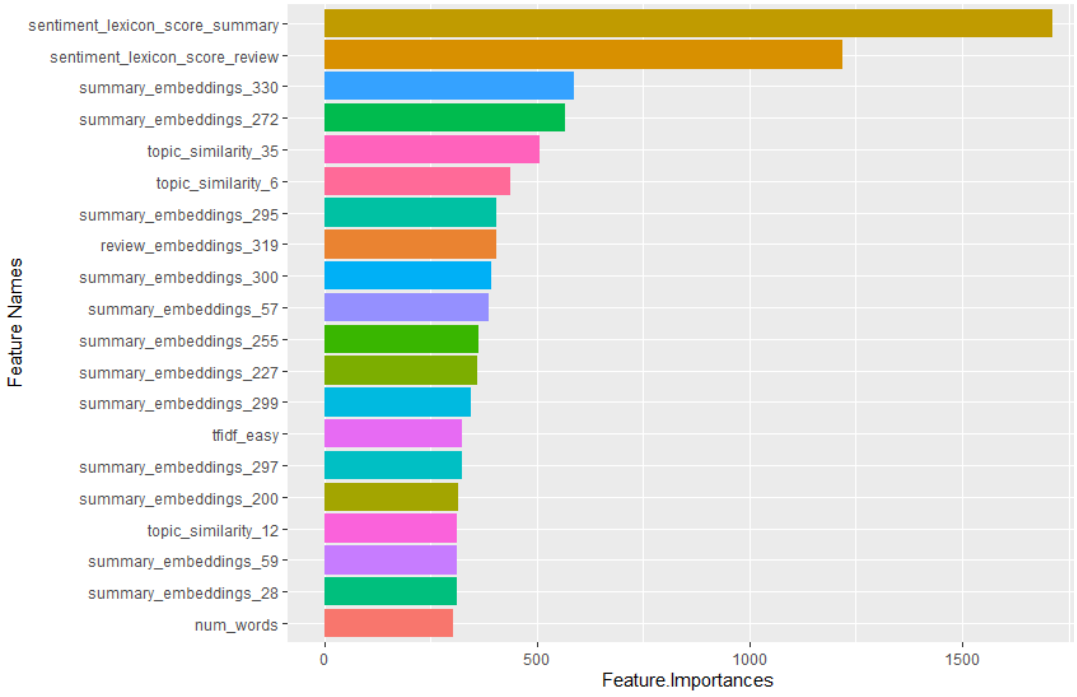

Figure 1 - Top 20 Most Important Features

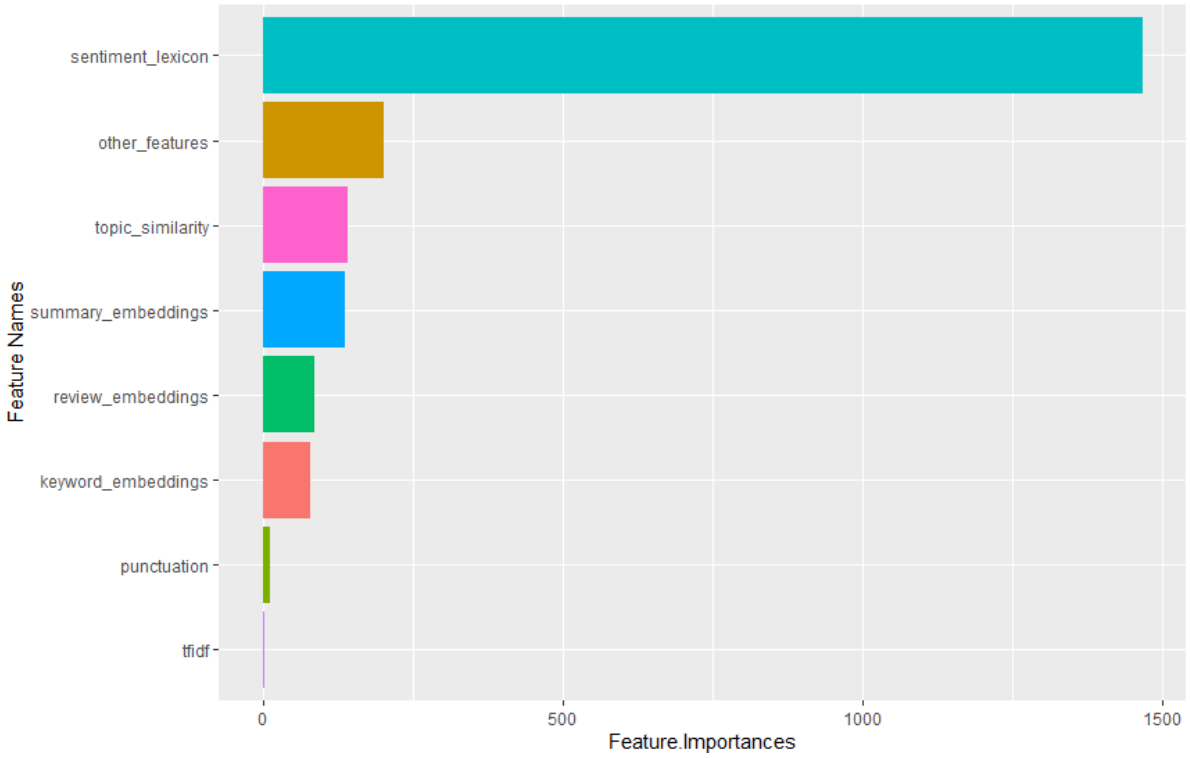

Figure 2 - Mean Feature Importance by Feature Types

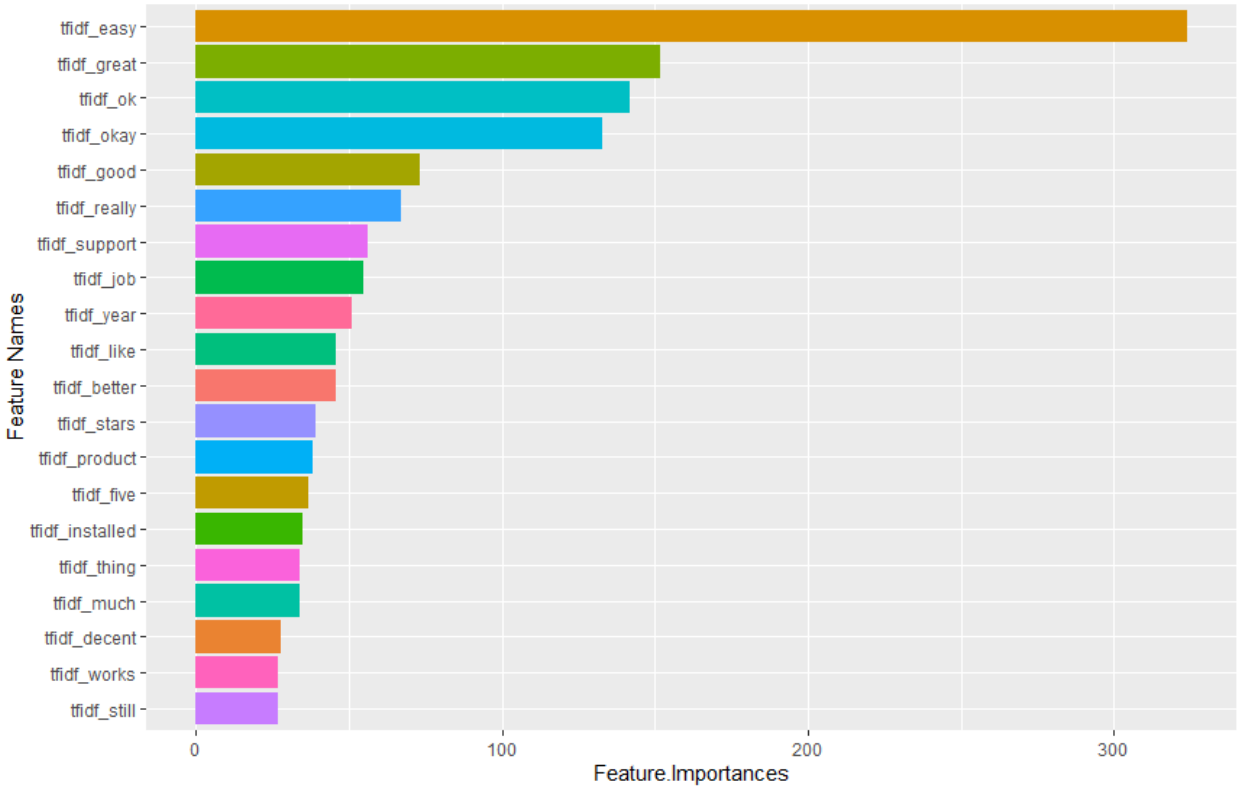

Figure 3 - Top 20 Most Important TFIDF word features

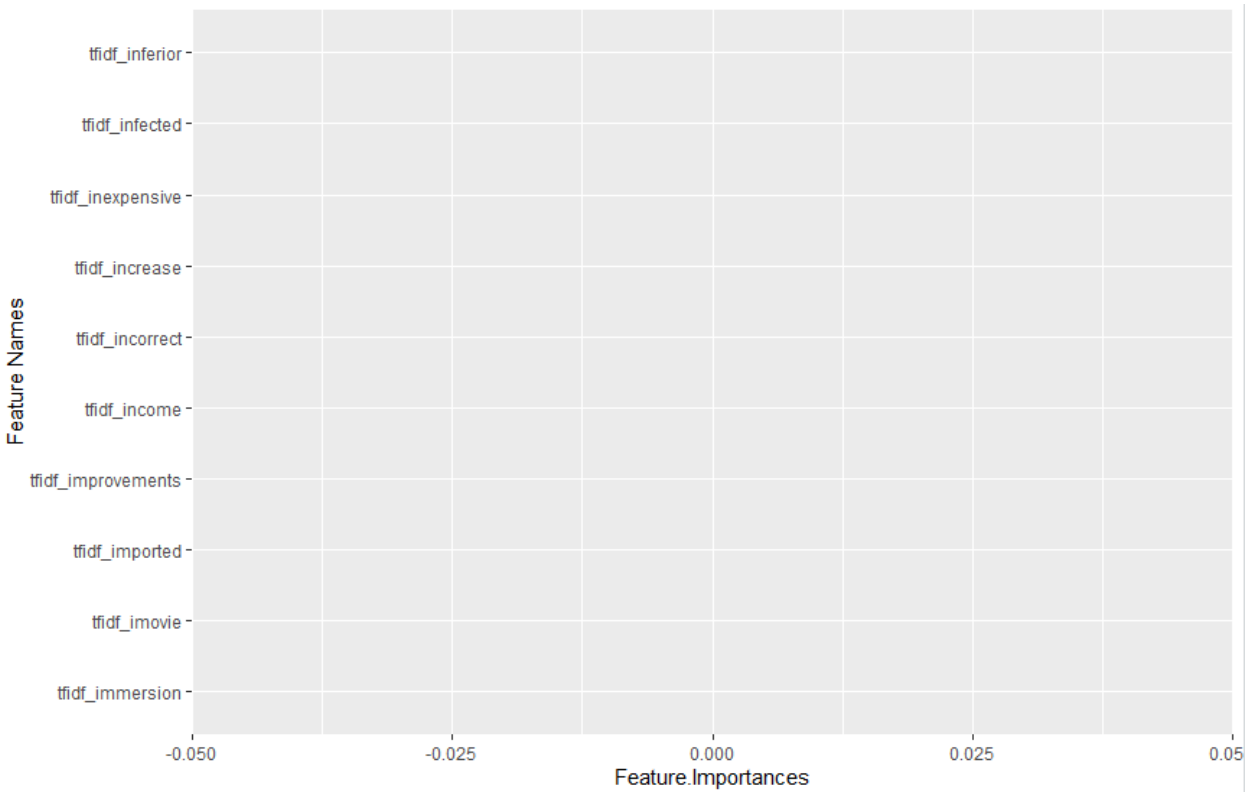

Figure 4 - Top 10 Least Important Features

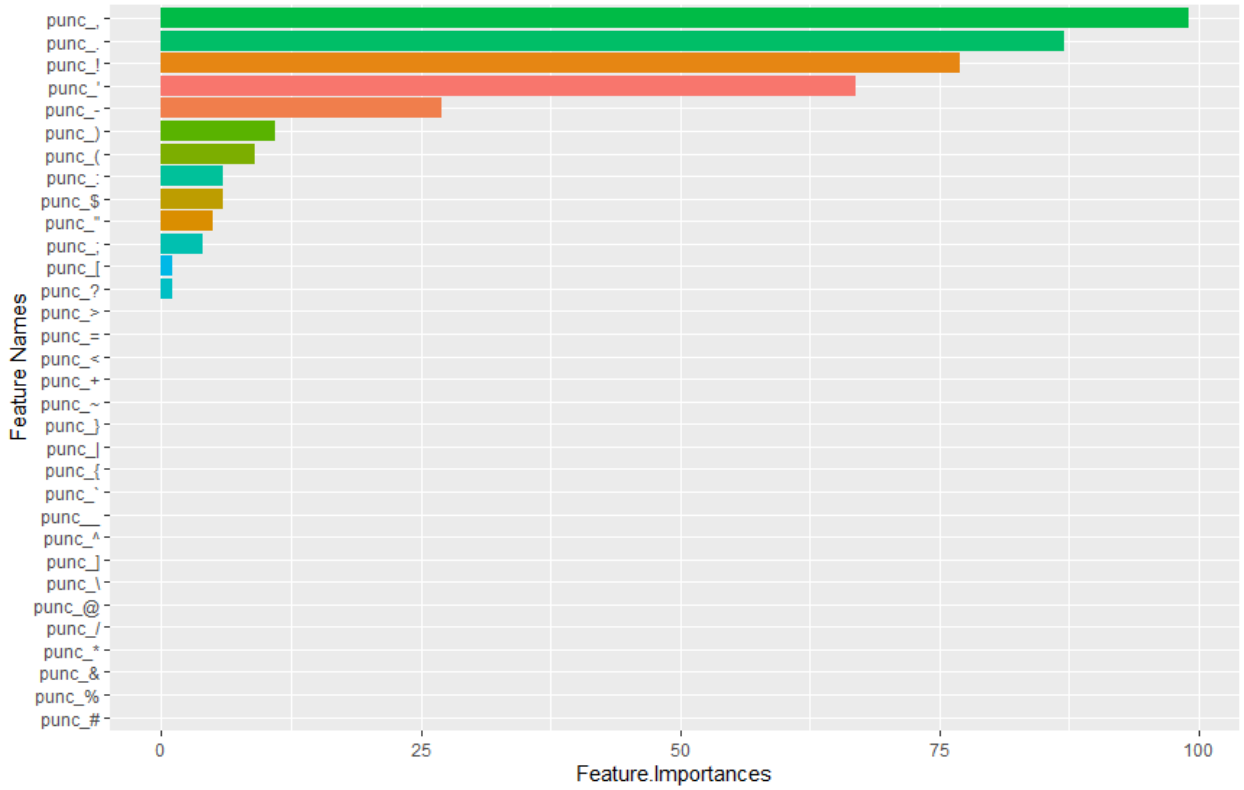

Figure 5 - Feature Importance of Punctuations

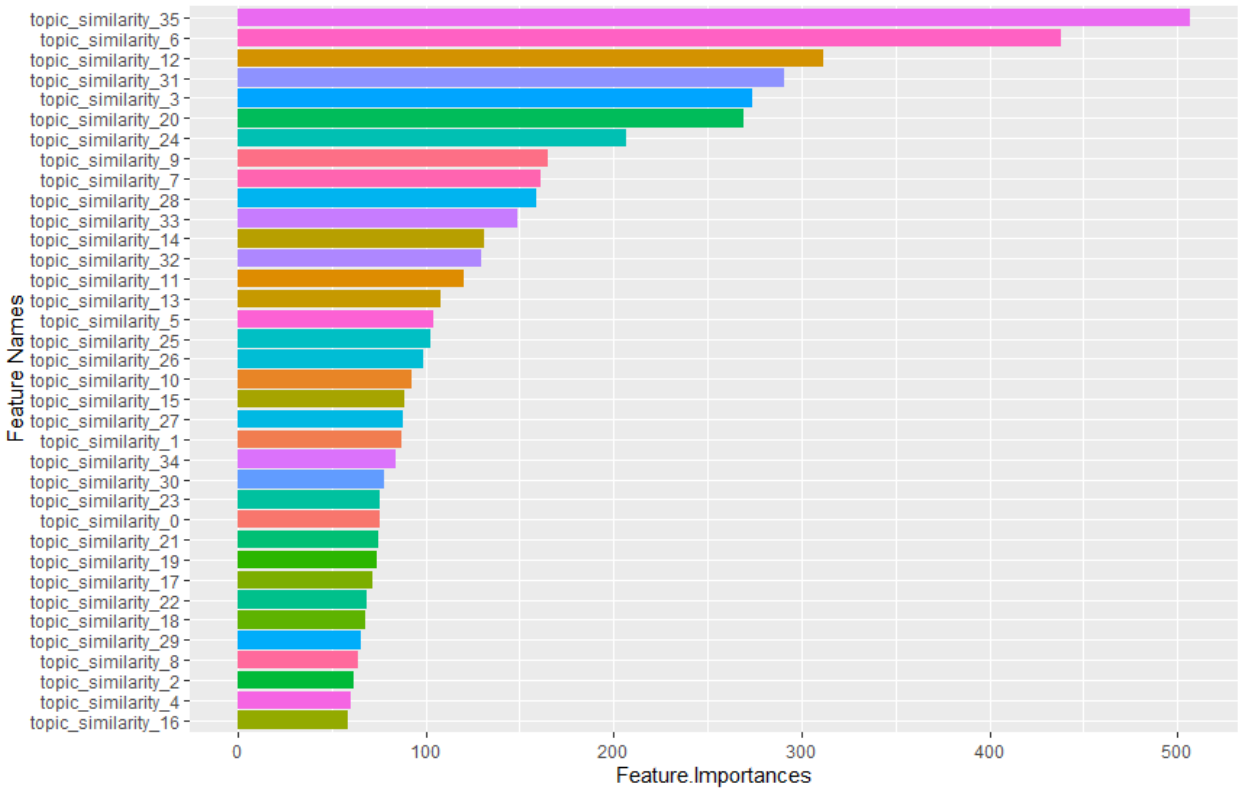

Figure 6 - Topic Similarity Feature Importance

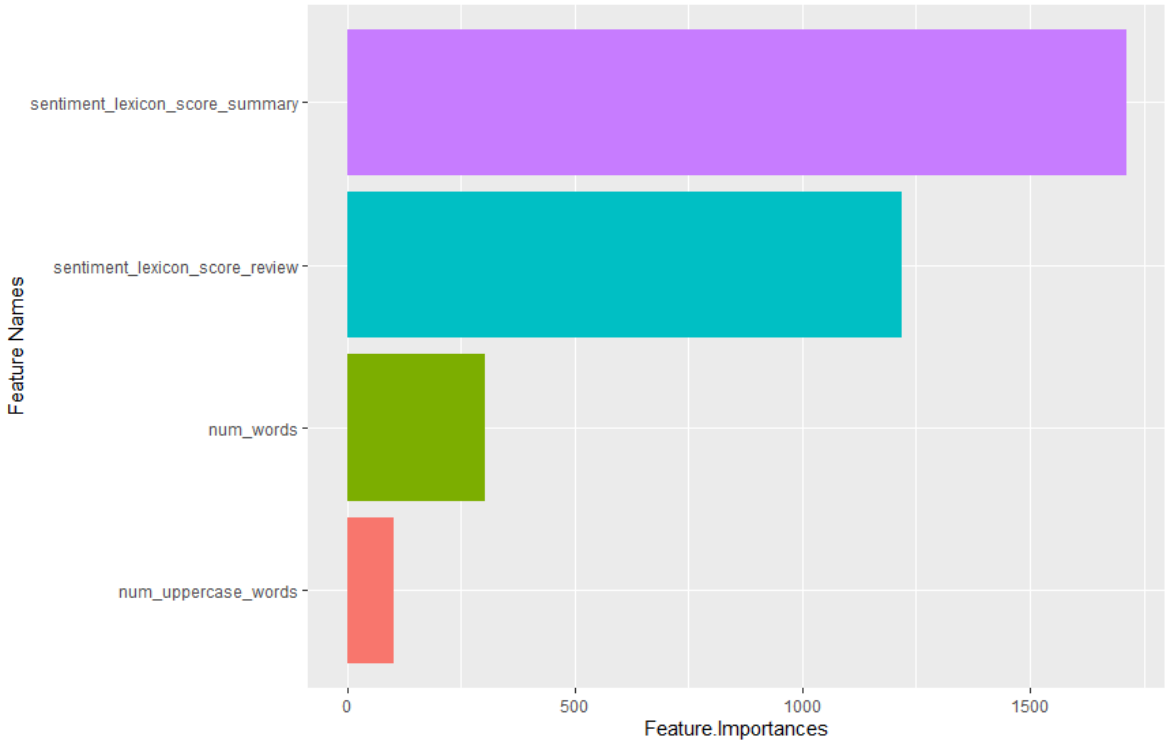

Figure 7 - Feature Importance - Sentiment Lexicon Scores and Engineered Features

Figure 8 - Topic 35 Word cloud

Figure 9 - Topic 6 Word cloud

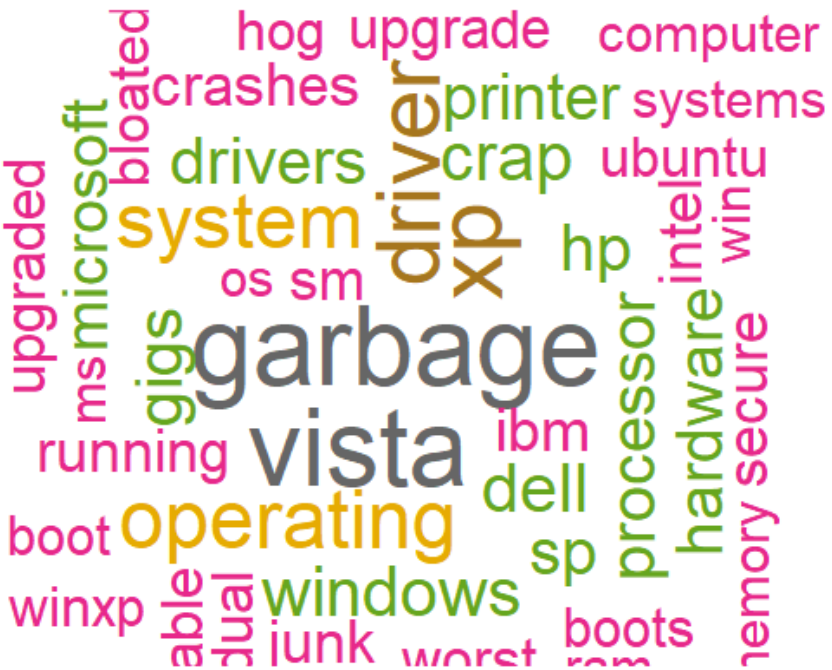

Figure 10 - Topic 12 Word Cloud

Figure 11 - Topic 2 Word Cloud

Figure 12 - Topic 4 Word Cloud

Figure 13 - Topic 16 Word Cloud

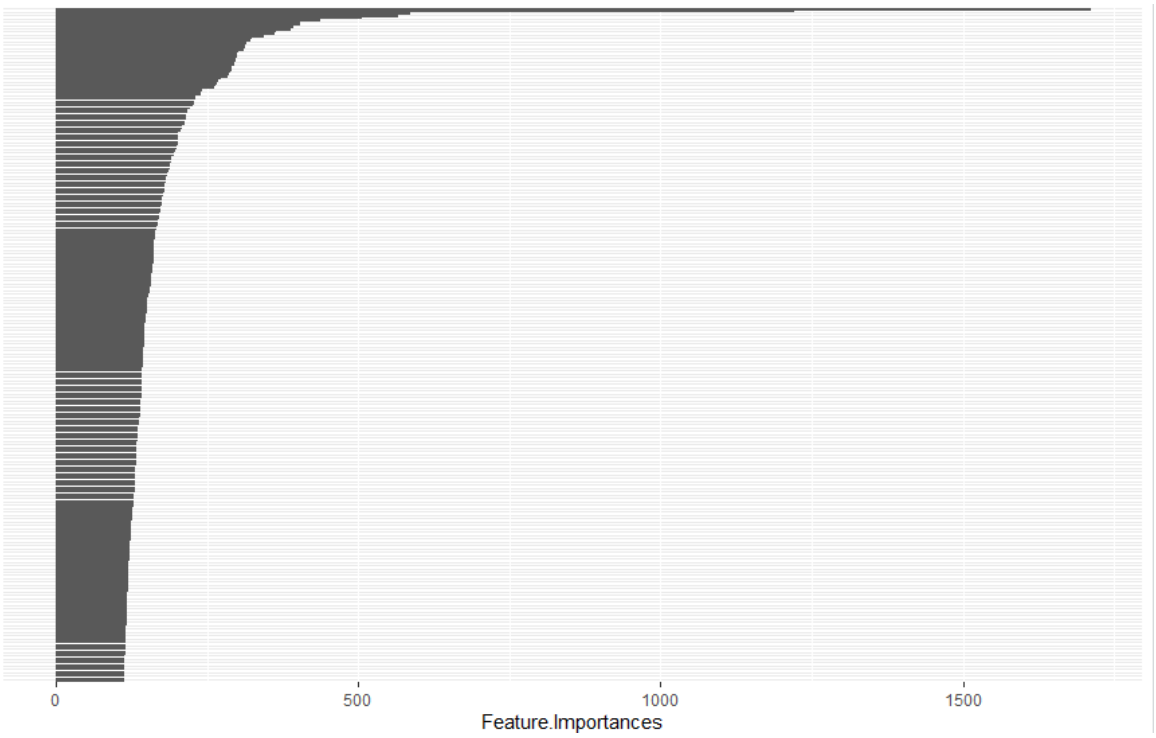

Figure 14 - Top 300 Features

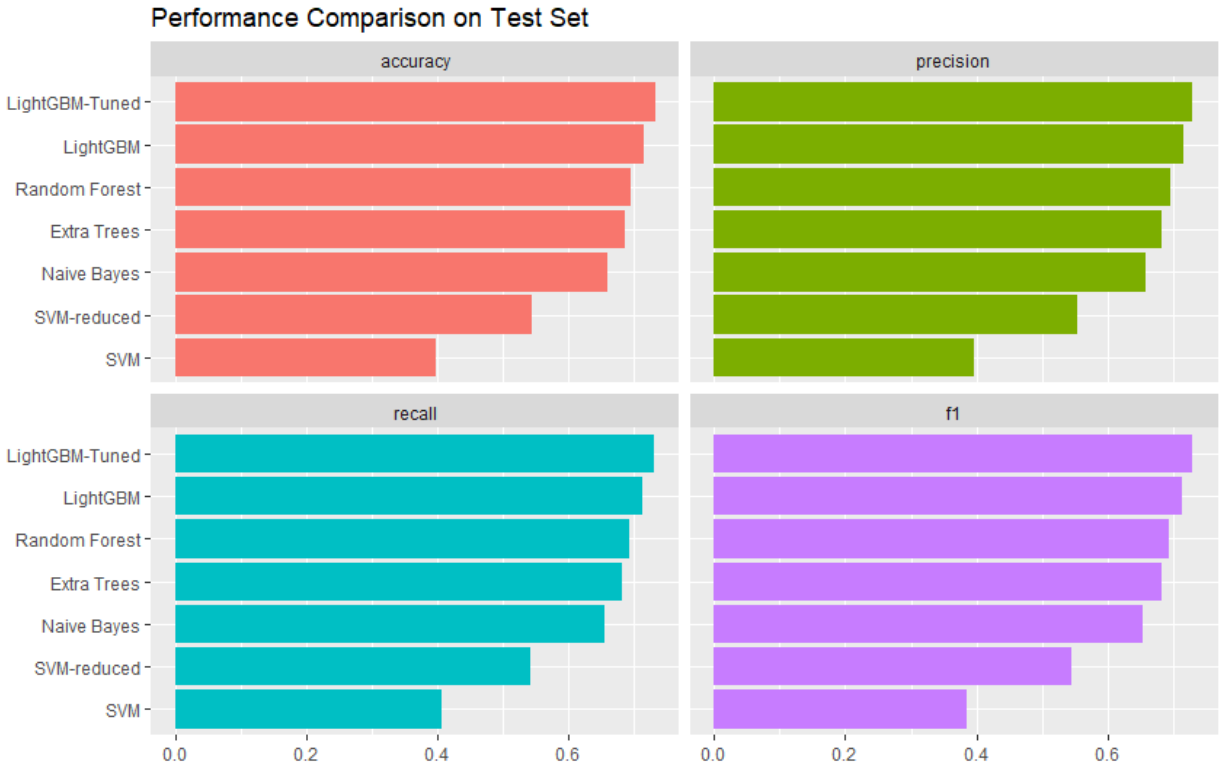

Figure 15 - Performance Comparison on Test Set

summary	reviewText	asin	brand	sentiment	reviewId
This is excellent software for those who want to use it as a "...	If you've been wanting to learn how to create your own web...	0321719816	Peach Pit Press	POS	1
Competent introduction to Dreamweaver and web principles.	I waited to complete the entire course before writing this re...	0321719816	Peach Pit Press	NEU	2
Learn Adobe Photoshop Lightroom 3 by Video (Learn by Vi...	As someone who has just upgraded from Lightroom version...	0321700945	Peach Pit Press	POS	3
For Highly Motivated and Patient People	There are over 100 video lessons here. Most users have give...	0321700945	Peach Pit Press	NEU	4
Good Intro to Flash CS5	This was the first Learn by Video series course that I've used,...	0321719824	Peach Pit Press	POS	5
A disappointment	Excel:Does not allow you to copy and paste sometimes. I lat...	0763855553	Microsoft	NEG	6
and easily corrupted. Outlook is all messed up	Been using Office for over twenty years. Still not worth the ...	0763855553	Microsoft	NEG	7
Suits my personal and small business needs	I have this running on my Macbook Air and two HP laptops....	0763855553	Microsoft	POS	8
Endless Updates	We got this for 2 Macs and 2 Windows machines. It works O...	0763855553	Microsoft	NEU	9
Five Stars	Has been very helpful.	0763855553	Microsoft	POS	10

Figure 16 - Sentiment Categorization Dataset


```
{'overall': 4.0,
 'summary': 'A solid overview of Dreamweaver CS5',
 'reviewText': "I've been using Dreamweaver (and it's predecessor Macromedia's UltraDev) for many years.  For someone who is an
experienced web designer, this course is a high-level review of the CS5 version of Dreamweaver, but it doesn't go into a great
enough level of detail to find it very useful.On the other hand, this is a great tool for someone who is a relative novice at w
eb design.  It starts off with a basic overview of HTML and continues through the concepts necessary to build a modern web sit
e.  Someone who goes through this course should exit with enough knowledge to create something that does what you want it do d
o...within reason.  Don't expect to go off and build an entire e-commerce system with only this class under your belt.It's impo
rtant to note that there's a long gap from site design to actual implementation.  This course teaches you how to implement a de
sign.  The user interface and overall user experience is a different subject that isn't covered here...it's possible to do a gr
eat implementation of an absolutely abysmal design.  I speak from experience.  :)As I said above, if you're a novice, a relativ
e newcomer or just an experienced web designer who wants a refresher course, this is a good way to do it.",
 'asin': '0321719816'}
```


Figure 17 - Raw Review Sample 1 (Extraneous Fields Removed)


```
{'brand': 'HOLT. RINEHART AND WINSTON', 'asin': '0030672120'}
```


Figure 18 - Raw Metadata Sample 1 (extraneous fields removed)

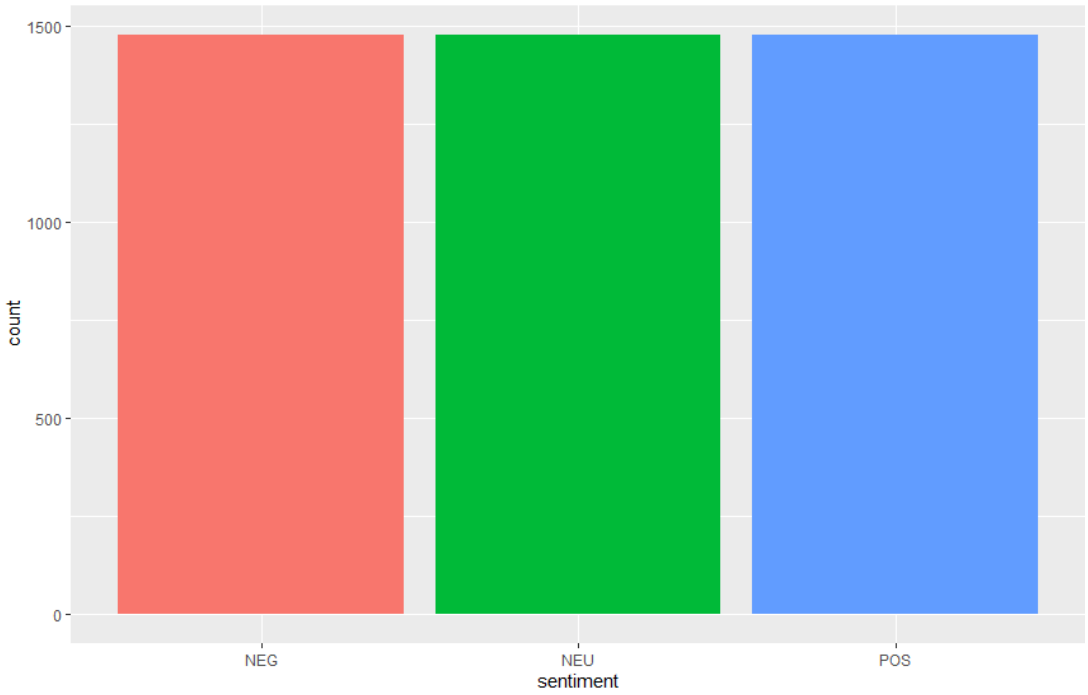

Figure 19 - Response Class Distribution

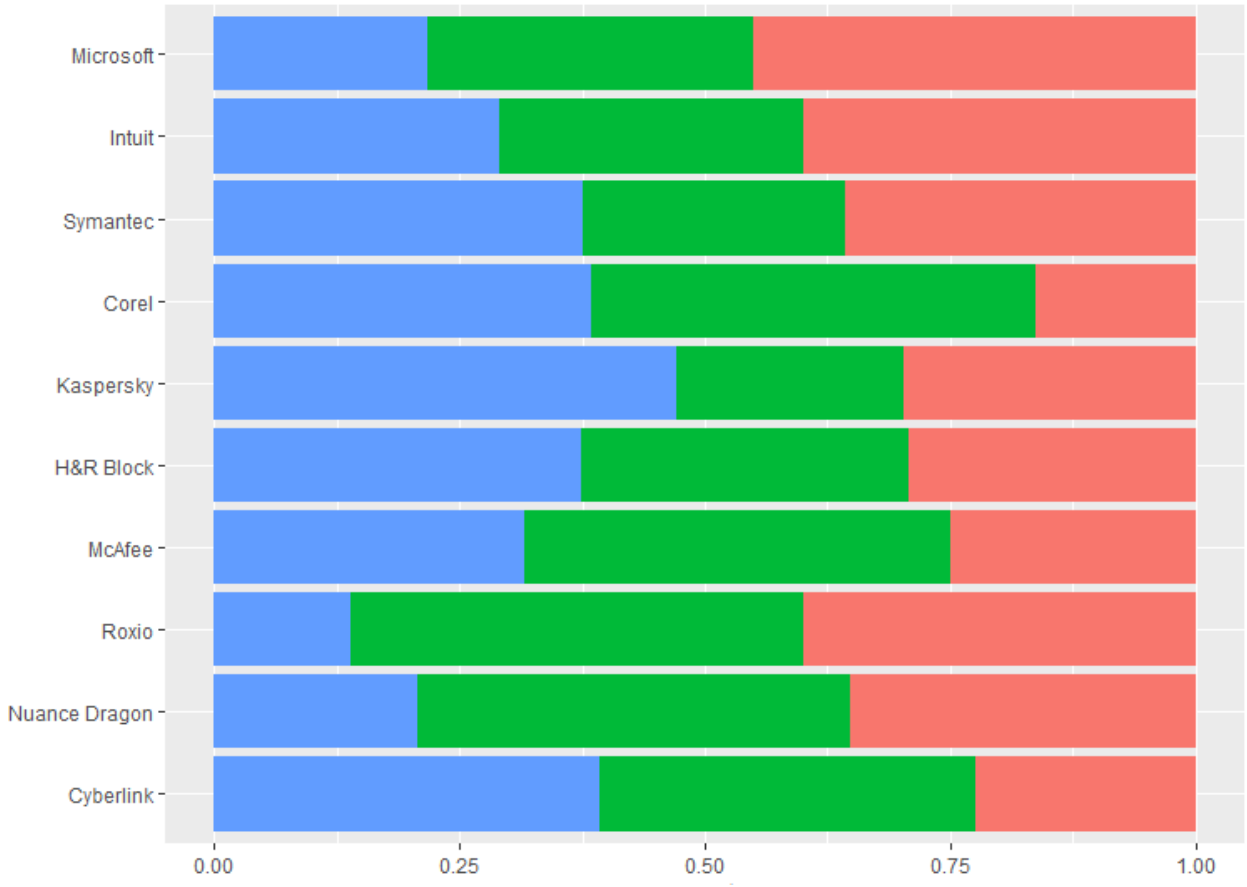

Figure 20 - Sentiment Distribution (Top 10 most frequent brands)

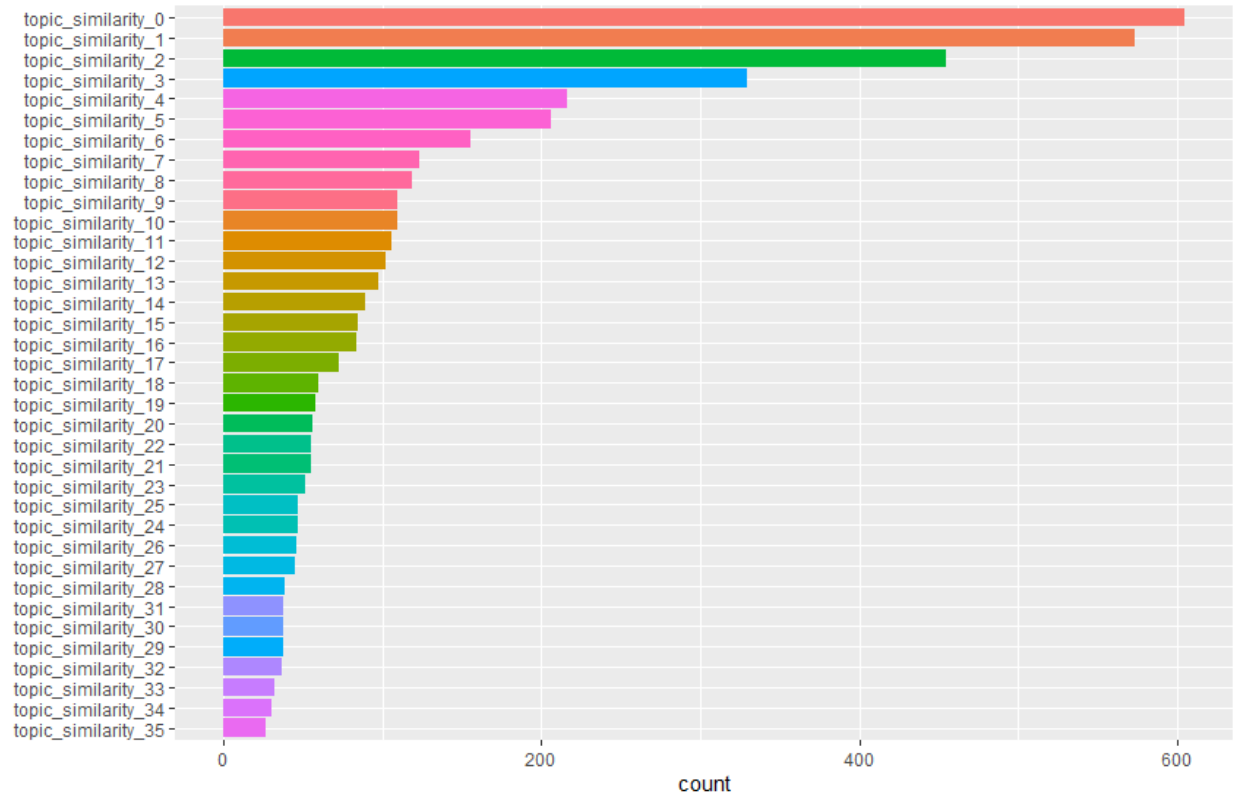

Figure 21 - Distribution of Topics


```
{'summary': 'Stay away from it',
 'reviewText': 'I program computers, and we had number of compatibility problems with this OS.  MS dropped much of good old DD
E, causing number of shell application to crash, including number of uninstall programs.',
 'asin': 'B00004W62O',
 'brand': 'Microsoft',
 'sentiment': 'NEG',
 'reviewId': 49,
 'response': -1}
```


Figure 22 - Processed Sample 49 (Most Similar Topic: 6)


```
{'summary': 'Excellent seller. definitely use again********',
 'reviewText': 'Finally getting what I was lookin for. Excellent seller. definitely use again********',
 'asin': 'B0014X5XEK',
 'brand': 'Apple',
 'sentiment': 'POS',
 'reviewId': 829,
 'response': 1}
```


Figure 23 - Processed Sample 828 (Most similar topic: 35)

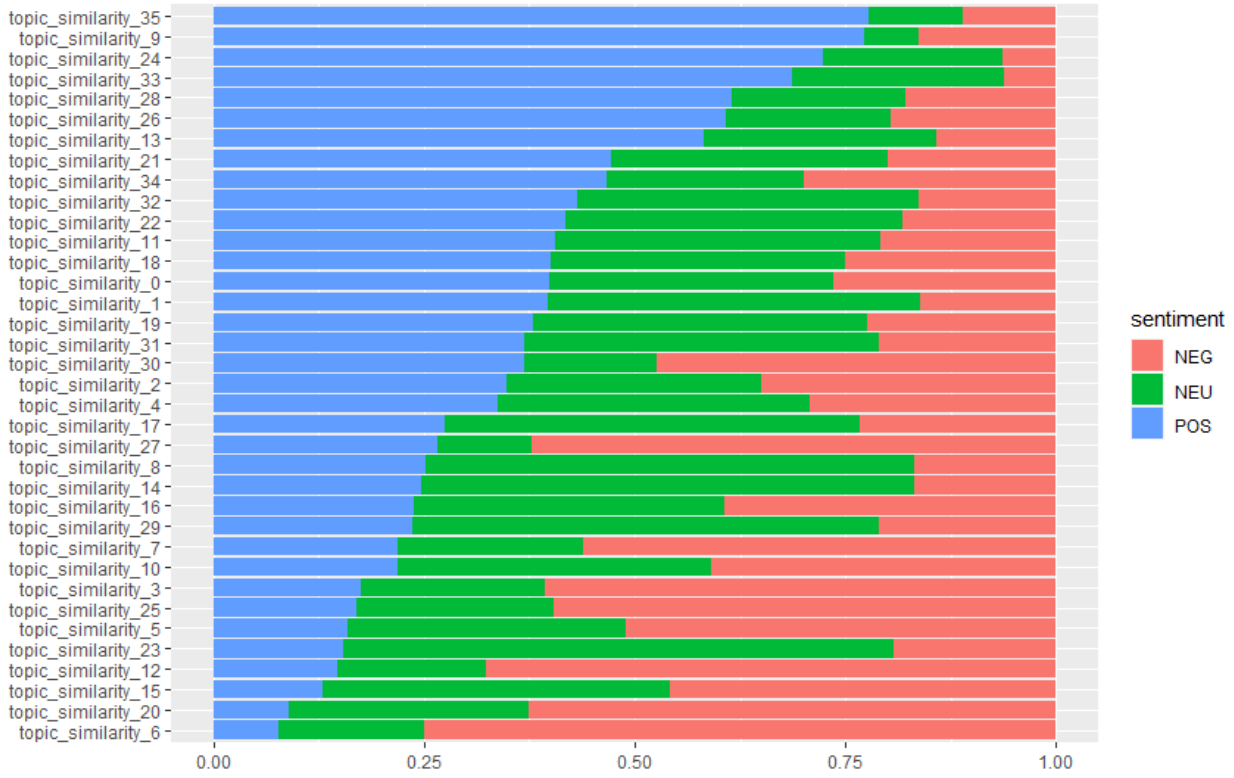

Figure 24 - Sentiment Distribution of Topics

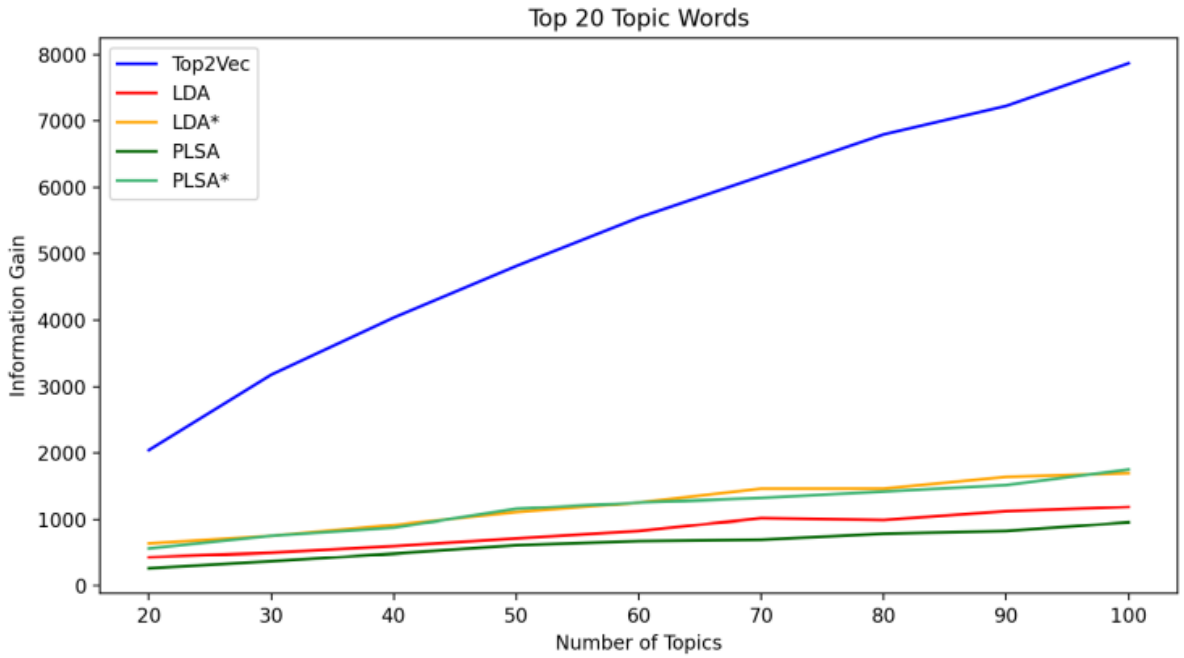

Figure 25 - Informativeness of Topic Words (Figure taken from Dimo Angelov [3])

	topic_similarity_0	topic_similarity_1	topic_similarity_2	topic_similarity_3	topic_similarity_4	topic_similarity_5	topic_similarity_6	topic_similarity_7
1	-0.073071726	0.17345826	0.01729862	-0.03697794	0.13899088	0.008364265	-0.00975300	-0.01094095
2	0.080383900	0.23847243	0.06823964	0.06213717	0.12862186	0.070838470	0.03145916	0.05821840
3	0.101315010	0.35587594	-0.01398280	0.07674670	0.13644351	0.037731014	0.08687694	0.10488292
4	0.103368160	0.30525672	0.13649435	0.01937337	0.16168060	0.071436040	0.02704033	0.03924098
5	-0.002345516	0.37359790	0.06698793	0.09742022	0.15858936	0.063632590	0.10242530	0.06740408
6	0.118894570	0.18593219	0.12822853	0.16691878	0.21778663	0.170919390	0.20130964	0.14323598
7	0.095472634	0.08186281	0.22306004	0.34968108	0.13132478	0.194454420	0.18194072	0.13538994
8	0.330298900	0.18025190	0.20778884	0.33427098	0.15036194	0.219495190	0.31277840	0.24182267
9	0.199806870	0.08185922	0.16582482	0.15828022	0.17120068	0.388414260	0.24172132	0.21824777
10	0.106312410	0.25903162	0.28712177	0.06443191	0.05644316	0.195886030	0.05242597	0.15361632

Figure 26 - Topic Similarity Features – similarity score of each review to each of the 36 topics

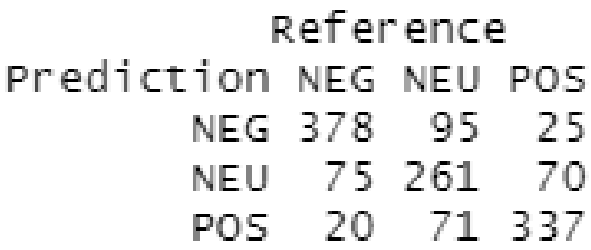

Figure 27 – Caret Confusion Matrix – LightGBM predictions

	Class: NEG	Class: NEU	Class: POS
Sensitivity	0.7992	0.6112	0.7801
Specificity	0.8603	0.8398	0.8989
Pos Pred Value	0.7590	0.6429	0.7874
Neg Pred Value	0.8861	0.8207	0.8949
Prevalence	0.3551	0.3206	0.3243
Detection Rate	0.2838	0.1959	0.2530
Detection Prevalence	0.3739	0.3048	0.3213
Balanced Accuracy	0.8297	0.7255	0.8395

Figure 28 - Caret Confusion Matrix Statistics - LightGBM predictions


```
{'summary': "I'm a bit leery",
 'reviewText': 'I\'m a bit leery of this product. I\'ve been a Norton customer since it first started providing security servic
es. The product detail says 25GB of Online Backup, but I receive a 5GB Online Back Up card via Amazon Vine. Where my skepticism
comes in is the product is an annual product like all other Norton products. In addition to this product being comparible to se
rvices like DropBox, Box, and SkyDrive, I looked through the TOS and the product paperwork that came in the package for what ha
ppens to my back ups provided I do NOT renew this service on an annual basis. Do I lose access to what I\'ve already backed up?
If this is the case, this could be detrimental and I would suffer a loss of my personal files. I currently have a ticket into N
orton to see what happens to my files provided I do not renew. I will post their response when I receive an answer, but I am no
t going to back up anything until I know the answer. One would thing that this would displayed clearly somewhere in the TOS. No
w the actual product is 25GB is a a good amount, however, it is still comparable to the other online storage and back up servic
es I mentioned above. With that being said, I already pay Norton in excess of $120 for Cyber Security products and multiple fam
ily back ups across devices, do I want to give them another $25 for online storage? You may want to bypass this one, but I will
post the response from Norton about the "shelf-life" of my back ups as soon as I hear from them.',
 'asin': 'B002X8V326',
 'brand': 'Symantec',
 'sentiment': 'NEU',
 'reviewId': 1379,
 'response': 0}
```


Figure 29 - Sample 1378 - Predicted Negative

{'summary': 'Easy download - good program',
 'reviewText': "Used H&R Block tax software for the first time last year - it was cheaper than what I had been using.Since they didn't constantly bug me to buy it again this year, and it worked well last year, I purchased again.I'm very glad that I'm not getting constant pleas to buy their software. And, it loaded last year's info just fine!I haven't efiled, yet - but that went well last year & I'm expecting it to go fine this year, too.Update: E-filing is great Having the name of tax professionals behind it is comforting But, downgraded to 3-stars because I had difficulty understanding some tax issues. Had to resort to the H&R block website (which was very good, and has helpful comments). But still, if links to IRS rules were available for each item I wouldn't have had to search as much.",
 'asin': 'B004A7Y0UK',
 'brand': 'H&R Block',
 'sentiment': 'NEU',
 'reviewId': 1622,
 'response': 0}

Figure 30 - Sample 1621 - Predicted Positive

{'summary': "WinZip is useful as it's ubiquitous, files can be unzipped by anyone, which makes it a good tool to manage and email large files. Of course, it's also useful to be able to zip them to store them in the Cloud, especially if you have the free versions and have much less storage space. It also allows you to carry multiple files on SD cards or USB drive or even your tablet.This version of WinZip connects directly to Dropbox, Google Drive etc so you can use it to organize all your cloud files at once. A password can be set for encrypted files for security.It also converts files to PDF, a useful feature for me, I have been using the free PrimoPDF so far. It can also create a read only PDF that is useful for things like resumes. I have only tried it with Microsoft office files so far.Overall, useful but also a little glitchy, if I did not have large files to compress I would keep using the free version.",
 'reviewText': "WinZip is useful as it's ubiquitous, files can be unzipped by anyone, which makes it a good tool to manage and email large files. Of course, it's also useful to be able to zip them to store them in the Cloud, especially if you have the free versions and have much less storage space. It also allows you to carry multiple files on SD cards or USB drive or even your tablet.This version of WinZip connects directly to Dropbox, Google Drive etc so you can use it to organize all your cloud files at once. A password can be set for encrypted files for security.It also converts files to PDF, a useful feature for me, I have been using the free PrimoPDF so far. It can also create a read only PDF that is useful for things like resumes. I have only tried it with Microsoft office files so far.Overall, useful but also a little glitchy, if I did not have large files to compress I would keep using the free version.",
 'asin': 'B00GDF84IG',
 'brand': 'Corel',
 'sentiment': 'NEU',
 'reviewId': 4404}

Figure 31 - Top result - querying documents

Figure 32 - predicting topic for new review

7 References:

Here, I append all the resources I utilized in this social media mining project.

[1] Data Source: Amazon review data (nijianmo.github.io)

[2] R package for top2vec: R: Distributed Representations of Topics (r-project.org)

[3] Top2vec research paper: [2008.09470] Top2Vec: Distributed Representations of Topics (arxiv.org)

[4] R package for generating TFIDFs: Term Frequency and Inverse Document Frequency (tf-idf) Using Tidy Data Principles (r-project.org)

[5] R package for generates embeddings with Transformers: RStudio AI Blog: State-of-the-art NLP models from R

[6] R package for keyword extraction: R: Keyword identification using Rapid Automatic Keyword... (r-project.org)

[7] Download of sentiment lexicons: WKWSCI Sentiment Lexicon v1.1 available for download | Chris Khoo (ntu.edu.sg)

[8] R package for training LightGBM Light Gradient Boosting Machine • lightgbm

[9] R package for training random forests, SVMs, Naïve bayes models: train function - RDocumentation

[10] R package for lemmatization: lemmatize_words: Lemmatize a Vector of Words in textstem: Tools for Stemming and Lemmatizing Text (rdrr.io)

END