Deep Text Classification in Social Media

Instagram Fashionista - Data Exploration

Kim Hammar, Shatha Jaradat

KTH Royal Institute of Technology

kimham@kth.se, shatha@kth.se

January 22, 2018

Introduction

We survey the basis for classifying the image contents of posts in Instagram based *solely on the associated natural language* in the form of captions, tags, and comments. In particular, we focus on the task of predicting fashion related characteristics of an image, such as type of clothing items, fashion brands of the clothes, material of the clothes, and general style description of the post. This analysis were performed on public Instagram data from a fashion-related user account called *hellofashionblog*[1]. The analysis is based on ≈ 3000 Instagram posts, associated with approximately 500000 user comments.

User analysis

Figures 1,3,5, displays the most common words in post tags, post captions, and post comments, respectively. Figures 2, 4, 6 contains wordcloud plots of the respective word frequencies. The statistics were collected by excluding stop words and tokenizing each post based on white space and semantic rules of the English language. As expected, the characteristics of the text is of a distinct nature compared to English language used in newspapers or literature. Evident from this analysis is that emojis (smileys) are very common, and that the frequency of jargon and words you'll not find in English dictionaries is unusually high compared to other sources of natural language.

Tags

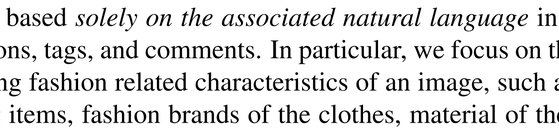


Figure 1: Histogram over the five most frequent tokens occurring in user tags made by the publisher of the Instagram posts.

Figure 2: Wordcloud over the most frequent user tags.

Captions

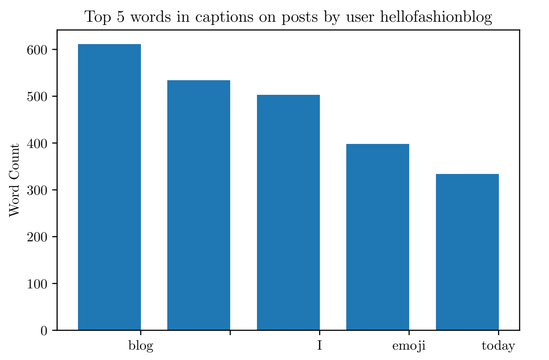
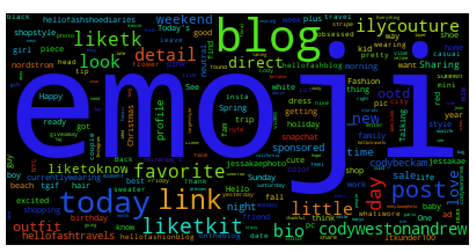


Figure 3: Histogram over the five most frequent tokens occurring in captions of the Instagram posts.



[1] https://www.instagram.com/hellofashionblog/

Figure 4: Wordcloud over the most frequent post captions.

Comments

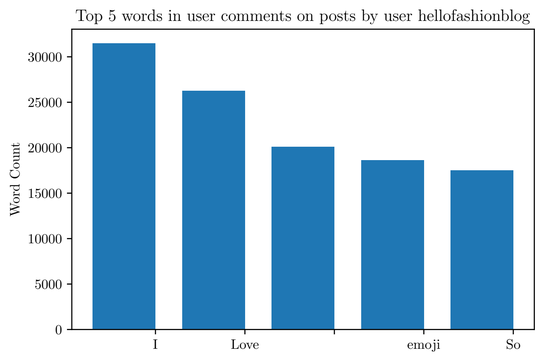


Figure 5: Histogram over the five most frequent tokens occurring in user comments on different Instagram posts.

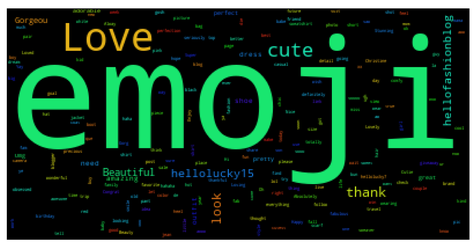


Figure 6: Wordcloud over the most frequent user comments.

Fashion Vocabulary

We have developed a fashion vocabulary to serve as a basis for unsupervised text mining of informative fashion descriptions embedded in natural language associated with Instagram posts. A subset of the vocabulary is shown in figure 7.

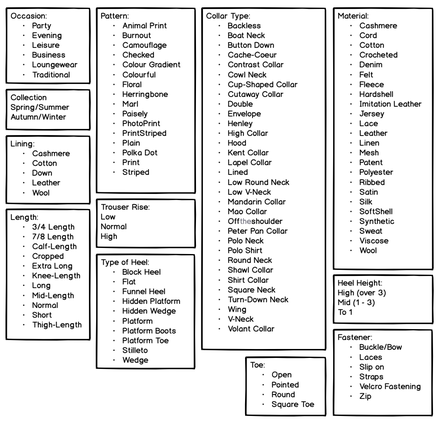


Figure 7: Subset of the fashion vocabulary.

Analysis of an Individual Post

In this section we analyze the natural language associated with a particular user post (Figure 8) and what type of predictive power is present in the textual data associated with the post.

Figure 8: A randomly selected post by the user account under analysis. The ground truth classification for this post would be something in the likes of *Items: dress, Brand: Bebe, Color: red, Material: polyester (not 100%), Style: looking up the dress in the retailer it is described as festive, sexy, valentine's day dress.*

To extract candidate fashion items, brands, materials, and style from the textual data associated with the post we did as follows.

1. Remove English stop words from the text.
2. Tokenize the text with an English parser.
3. Fetch word vectors that have been pre-trained on a large English corpus (trained with word2vec).
4. Convert each token into its corresponding word vector.
5. Tokenize a small fashion vocabulary of items, brands, materials, and styles.
6. Convert each token from the fashion vocabulary to its corresponding word vector.
7. Convert each token into its corresponding lemma.
8. For each word associated with the post, compute the cosine similarity with all the words in the fashion vocabulary and extract all words with a similarity score > 0.7 into either a candidate clothing item, style, or material, based on the type of word it matched to.
9. In our vocabulary of fashion brands we have some companies named after common English words such as "Hope", "Love" etc, making semantic matching difficult. Therefore, brands are extracted by exact syntactic matching instead.

Figures 9-11 shows the candidate fashion items, fashion styles, and item materials. For this post there was no match with any fashion brand contained in our vocabulary. After manual inspection of the Instagram post, it was discovered that the brand of the dress was referenced in the textual data, but the mentioned brand (Bebe) were not present in our vocabulary and thus did not yield a match. Figure 12 displays the brand candidate extracted from another random post that yielded a match in our vocabulary.

The results indicate that the text associated with the image is rich on clues of the style details of the image. Especially the item extraction was effective, generating the correct ground truth class. The inference of styles of the image were not perfect but not bad either, it managed to extract "elegant", "chic" and "sexy" which are correct classifications, while also extracting "casual" which is clearly wrong. The brand and material inference were less successful. The failure to extract the fashion brand can be deduced to the fact that the brand were missing from the vocabulary of fashion brands that were used. The analysis extracted "Felt" as a probable material of the dress, which is not correct. In this case we believe that the difficulty of extracting the correct material is due to the fact that it were not mentioned in the text, and "Felt" was extracted because it is a homonym word, "Felt" can refer to other meanings than the material, depending on the context of the usage of the word.

Candidate Items

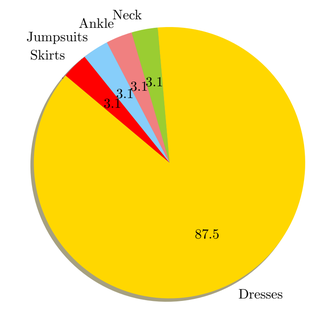


Figure 9: A pie plot highlighting the relative frequency of words with a high semantic similarity with fashion items in our vocabulary.

Candidate Styles

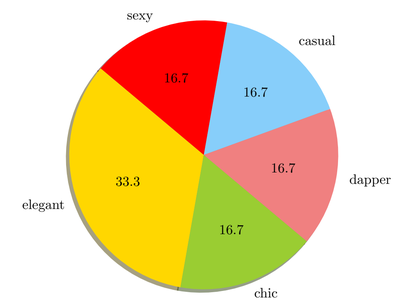


Figure 10: A pie plot highlighting the relative frequency of words with a high semantic similarity with style descriptions in our vocabulary.

Candidate Materials

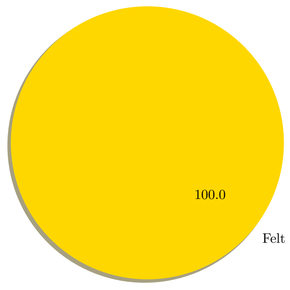


Figure 11: A pie plot highlighting the relative frequency of words with a high semantic similarity with clothing materials in our vocabulary.

Candidate Brands

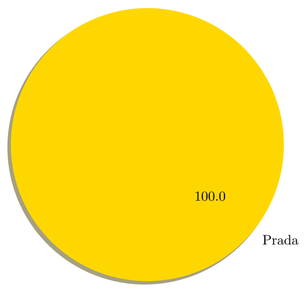


Figure 12: A pie plot highlighting the relative frequency of words with a high semantic similarity with fashion brands in our vocabulary.

Post coverage by Our Fashion Vocabulary

In our interest is that the fashion vocabulary has a high coverage over the posts, ideally extracting candidate items, materials, brands, and styles for each post that is fashion related. Figures 13 - 17 shows the coverage over the entire set of posts in finding candidate items, styles, brands, and materials, respectively. As our data exploration is unsupervised, the true number of fashion-related items posted by this user is unknown. To summarize the results, around $600/3000$ posts had no item candidate discovered, $100/3000$ posts had to style information discovered, $2200/3000$ had no brand discovered, and $1200/3000$ posts had no clothing-material discovered. The results indicate that our vocabulary of brands had a poor coverage over the set of posts. This is likely a combination of missing words in our vocabulary, that the brand is not always explicit in the textual data associated with a post, and that our information extraction techniques for discovering brands in the text (syntactic matching) is not perfect.

Notable is that a few of the histograms demonstrate the long-tail phenomena, e.g in figure 13 the majority of posts are within the range of $0 - 50$ candidate items, yet there are posts in the whole range up to 250 candidate items. The long-tail shape of the distribution indicates a power-law distribution, as evident in the log-log plot shown in figure 14.

Fashion Item Coverage

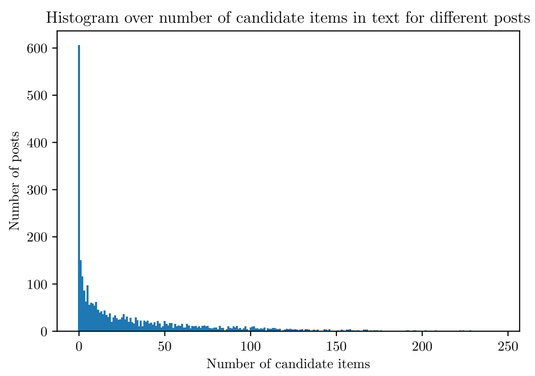


Figure 13: Histogram showing the distribution of the number of candidate items identified for different posts.

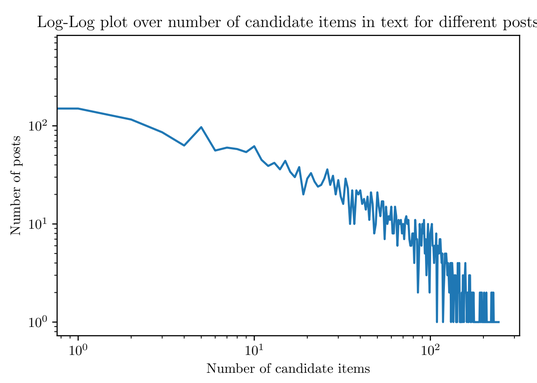


Figure 14: Log-log plot showing the distribution of the number of candidate items identified for different posts.

Fashion Style Coverage

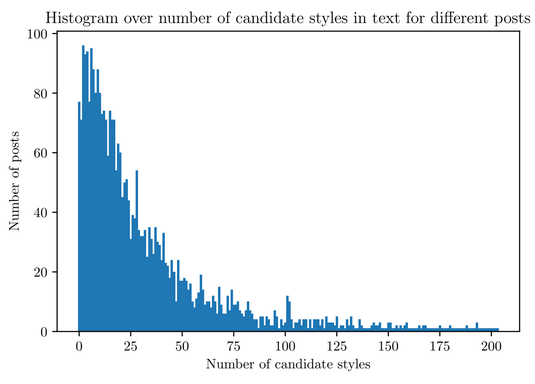


Figure 15: Histogram showing the distribution of the number of candidate styles identified for different posts.

Fashion Brand Coverage

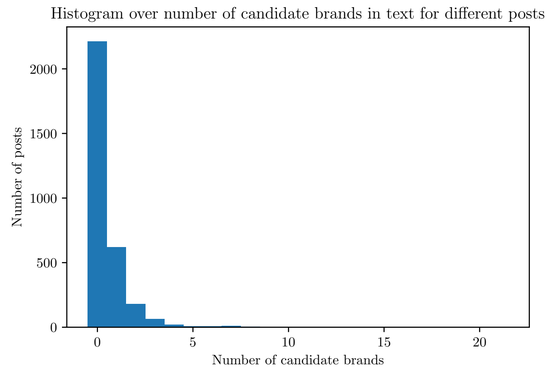


Figure 16: Histogram showing the distribution of the number of candidate brands identified for different posts.

Item Material Coverage

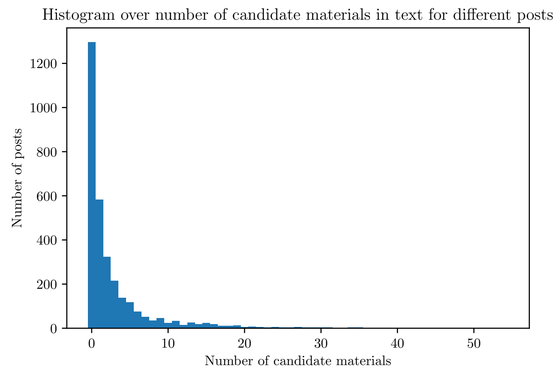


Figure 17: Histogram showing the distribution of the number of candidate materials identified for different posts.

A community of fashion related Instagram accounts are part of the LiketoKnow.it influencer network. Users that are partnered with LiketoKnow.it annotate their posts on Instagram with extra clothing details, that then are published and aggregated by the LiketoKnow.it service. Enabling followers of the influencers to look up the fashion items of a post in their corresponding shopping websites, under the assumption that the influencer had annotated the post with enough information. The Instagram account we focused on in this study are part of the LiketoKnow.it influencer network. Figure 18 shows the fraction of posts where we were able to find a LiketoKnow.it link by matching with regular expressions. Only 5% (≈ 150 posts) of the posts had the extra annotations and a LiketoKnow.it link, far less than the number of fashion related images discovered by our text mining techniques.

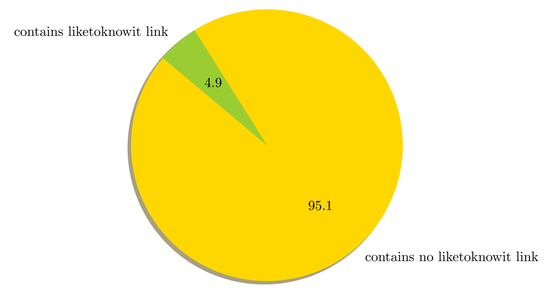


Figure 18: Pie plot showing the fraction of posts that contains a liketoknowit link.

Learning Vector Representation of Words

Semantic Word Vector representations serve as the foundation for the unsupervised text mining techniques outlined earlier in this document. In particular, word vectors that are able to generalize between different fashion words and concepts are desirable. In the results described in this document, word vectors pre-trained with word2vec on an English corpus based on Wikipedia was used. The advantage of the pre-trained vectors are that they contain a huge vocabulary of generic English words. However, an important disadvantage with these vectors are that they don't include social media jargon such as hash-tags or emojis in the vocabulary. In light of this disadvantage, we trained word vectors on a corpus of all the words occurring across all posts of the user account under analysis. In summary, the vectors trained on social media text are able to better capture the jargon of social media, including vector representations for emojis and hash-tags. A comparison of how well the two vector representations are able to capture similarities between common fashion words are outlined in table 1. The vectors trained on the social media corpus shows promising results bearing in mind that the corpus is so small in comparison with Wikipedia corpus, indicating that training word vectors on a larger social media corpus (e.g 100 users) might give even better results.

First Word	Second Word	Vectors trained on Wikipedia ($\mathcal{V}_1$)	Vectors trained on Instagram posts ($\mathcal{V}_2$)
gucci	fashion	0.47	0.14
#gucci	gucci	0	0.85
gucci	prada	0.82	0.76
sweater	shirt	0.72	0.89
sandal	shoe	0.66	0.35
coat	jacket	0.68	0.94
jacket	top	0.35	0.92
blouse	top	0.31	0.81
👛	purse	0	0.34
👖	jeans	0	0.41
👖	dress	0	0.08
👗	dress	0	0.31
👒	hat	0	0.45
👜	handbag	0	0.62
👜	dress	0	0.11
👙	bikini	0	0.60
👢	boots	0	0.23
👠	heels	0	0.48
👠	shoe	0	0.57
👟	sneaker	0	0.83
😎	sunglasses	0	0.38
🕶	sunglasses	0	0.42
👔	tie	0	0.42

Table 1: Cosine similarity between fashion words/tokens using two different semantic vector representations $\mathcal{V}_1$ and $\mathcal{V}_2$. $\mathcal{V}_1$ are vectors trained on a generic English corpus based on Wikipedia posts. $\mathcal{V}_2$ are vectors trained on 3000 social media posts.

Figure 19 and Figure 20 shows some examples of word vectors learned on the social media corpus. Before plotting, the vectors were reduced to two dimensions. Noteworthy is that the social media corpus contains a lot of user-tags of the form *@username*, which probably should be filtered out prior to training the word vectors as they bare no predictive power for the task of fashion classification.

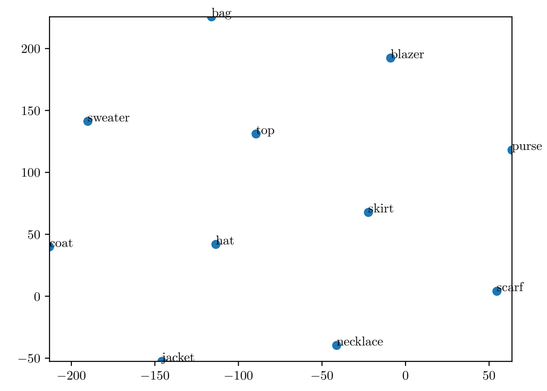


Figure 19: Word vector representation of the words closest to the word "jacket". The vectors were trained on the social media corpus. The original vectors vectors live in 100 dimensions, and were reduced to a two dimensional euclidean space using T-SNE.

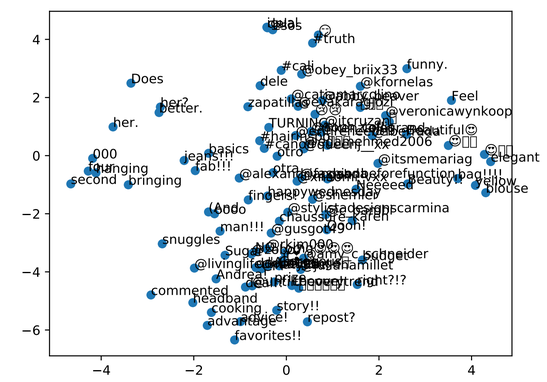


Figure 20: Word vector representations of 100 random words from the social media corpus. The original vectors vectors live in 100 dimensions, and were reduced to a two dimensional euclidean space using T-SNE.

Word2Vec, Glove, FastText There are three different approaches for computing word vectors, Word2Vec, Glove, and FastText, developed by Google, Stanford, and Facebook respectively. In these experiments we used Word2Vec simply because it is the most mature and accessible method of the three. However, many independent experiments demonstrate that Glove tend to out-perform Word2Vec, making it a better choice for future experiments. FastText is specialized for languages rich on morphology, working on character level predictions rather than word level like Glove, and Word2Vec. Our dataset is mostly based on English, which is not a language with high morphology in relation to other languages, discouraging the usage of FastText. However, our linguistic knowledge is too ground to judge if FastText would be suited for our corpus or not. The most safe approach is probably to compare all three techniques if time and resources allows.

Collective Style Analysis of User

Apart from classifying fashion details of a single posts, we are also interested in combining post classifications of a user to determine the overall style. Figures 21 - 28 shows summary statistics of the most common items, style, materials, and brands that were identified across all posts by this user.

Notable in these results is that "Felt" were identified as a common clothing material, I suspect that many of the occurrences of "Felt" refer to something else than the clothing material. Additionally, "Hope" and "else" were identified as two common clothing brands, which might be misleading as those words also have different meanings.

Clothing Items

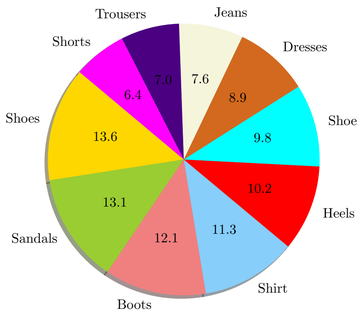


Figure 21: Pie plot showing the ten most common clothing items by this user.

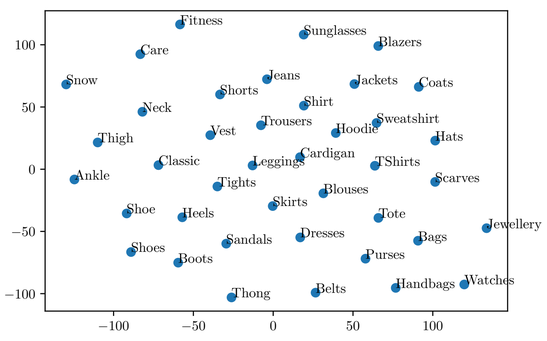


Figure 22: A scatter plot of the word vectors corresponding to the 40 most common clothing items of this user. Similar clothing items should appear close in the euclidean space.

Clothing Style

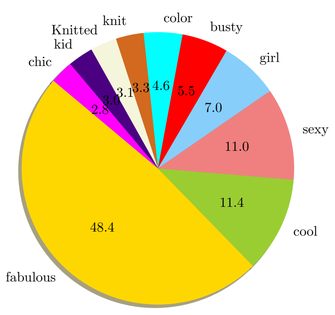


Figure 23: Pie plot showing the ten most common style descriptions of this user.

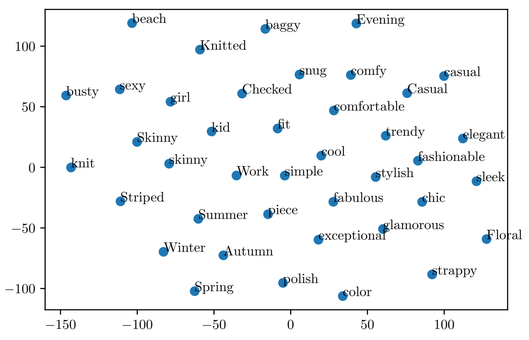


Figure 24: A scatter plot of the word vectors corresponding to the 40 most common clothing style descriptions of this user. Similar clothing styles should appear close in the euclidean space.

Clothing Materials

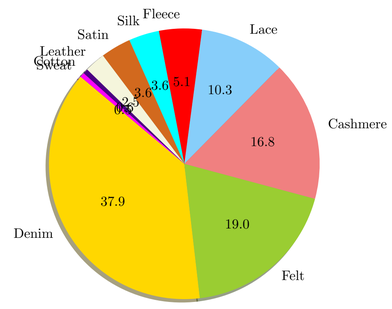


Figure 25: Pie plot showing the ten most common clothing materials worn by this user

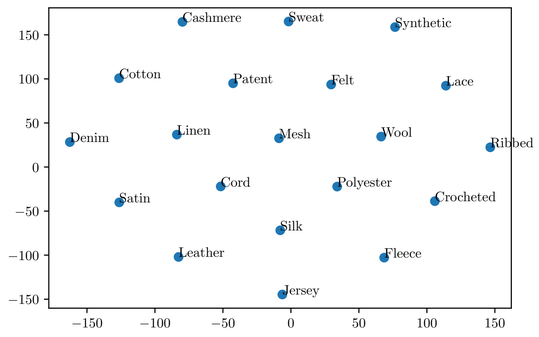


Figure 26: A scatter plot of the word vectors corresponding to the 40 most common clothing materials of this user. Similar clothing materials should appear close in the euclidean space.

Clothing Brands

Figure 27: Pie plot showing the ten most common clothing brands worn by this user.

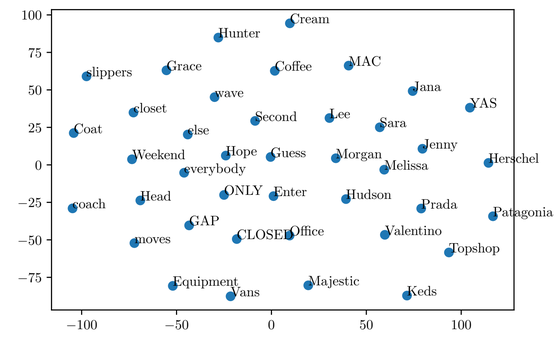


Figure 28: A scatter plot of the word vectors corresponding to the 40 most common clothing brands of this user. Similar clothing brands should appear close in the euclidean space.

Conclusions

The predictive power for fashion classification present in the textual data associated with social media posts is high. The social media domain poses special challenges by the unique vocabulary of words used and the noisy characteristic of the text. Generic word vectors trained on an English corpus are not able to capture the characteristics of the social media domain. Therefore, word vectors trained on a social media corpus is an attractive alternative.

Forthcoming Research

The information extraction techniques performed in this document can be used as labeling guidelines for crowdsource labeling by for instance the Amazon Turk service. Additionally, it can serve as weak-supervision signals or feature extraction in machine learning pipelines. The text mining techniques are limited in that they are static and hard-coded, ideally we seek to learn more *general* patterns for extracting fashion characteristics of social media posts. In consideration of this, we will look at applying a combination of deep learning and ontology interpretation of natural language. Finally, there are room for improvements regarding the text mining, such as extracting links from the text that could potentially point to fashion retailers, using a knowledge graph or semantic web API to look-up brand names dynamically, looking up sets of synonyms from WordNet, perform more sophisticated filtering of the text and removing user tags, and extending the fashion vocabulary. Additionally, considering the positive results of training word vectors on this relatively small corpus of social media posts, we will consider using word vectors trained on a larger corpus of social media posts to replace the vectors trained on English Wikipedia that were used in the information extraction process outlined in this document. Specifically, we might consider hand-labeling similarity scores between fashion related words and using that as an evaluation set to evaluate our trained word vectors against vectors pretrained on a Wikipedia corpus.