The importance of brand affinity in luxury fashion recommendations

DIOGO GONCALVES, Farfetch
LIWEI LIU, Farfetch
JOÃO SÁ, Farfetch
TIAGO OTTO, Farfetch
ANA MAGALHÃES, Farfetch
PAULA BROCHADO, Farfetch

Recommender systems in the context of luxury fashion need to have expert domain knowledge to offer the informed experience expected by the customers of this sector. Fashion experts have a strong understanding of the intricacies of the fashion scope. The brands and designers are some of the most important features of this landscape and the affinity between them is not always easy to grasp. This paper proposes an application of state-of-the-art NLP techniques to map the knowledge provided by experts in the form of texts. The outcome was a process to extract brand embeddings which mirror the semantic adjacency between all the brands in a catalogue. To test the utility of such an approach, we conducted extensive offline and online tests which have proven the positive reaction of the customers to the new feature. We applied the embeddings as boosting to a base recommender system and we observed an engagement uplift of up to 10%, and applied the embeddings as a content-based recommender to obtain an engagement uplift of up to 3%. Overall, we are confident of the importance of brand affinity information in recommender systems in the luxury fashion domain.

CCS Concepts: • **Information systems** → **Recommender systems**.

Additional Key Words and Phrases: Recommender Systems, Luxury Fashion, word embeddings, NLP, Farfetch

ACM Reference Format:
Diogo Goncalves, Liwei Liu, João Sá, Tiago Otto, Ana Magalhães, and Paula Brochado. 2020. The importance of brand affinity in luxury fashion recommendations. In *Proceedings of 2nd International Workshop on Recommender Systems in Fashion, 14th ACM Conference on Recommender Systems (fashionXrecsys'20).* ACM, New York, NY, USA, 13 pages.

1 INTRODUCTION

Across the whole fashion e-commerce sector, customers should experience more and more recommendation systems to be tailored to their needs and fashion tastes [5]. In this luxury fashion context, and particularly at Farfetch, customers expect not only a personalised experience, but also an informed opinion regarding latest trends, new arrivals and fashion understanding. Farfetch is the leading platform for online luxury fashion shopping. We count with the biggest catalogue of luxury items in the World with more than 3 million products and more than 10 thousand brands and high-end designers. Moreover, we sell products worldwide to more than 2 million customers. Our customers expect the highest standards regarding shopping and they usually pursue great experiences in their journeys as clients.

Manuscript submitted to ACM

Luxury fashion is a form of art, and that has to be taken into consideration whenever we are recommending a product to a possible customer. Experiences and expectations in luxury fashion might be closer to music and *beaux-arts* than to fast fashion shopping. Designers are artists and the work they do in their brands lead to legions of fans to follow them closely and rejecting the idea of art directors moving between brands [12]. Therefore, it's paramount that the recommender systems in such a domain have the strongest fashion understanding of what a brand means on a global and on a personalised scale.

In this context, pure collaborative filtering (CF) approaches are expected to fail due to the behavior of the luxury customer who wants exclusivity. There are many products in a luxury catalogue that contain very few items in stock, being sold-out after just one purchase. This particularity causes products to no longer be available at the time a CF algorithm identifies them as good recommendations to a user. This scenario leads to the need for content-based hybrids, which leverage the users and items information to deliver the level of personalisation needed by a recommender system in the fashion context. There are many ways to incorporate content in recommender systems to create hybrid solutions. However, the quality of the outcome is very dependent on the quality of the provided information.

Our proposal is to implement a neural network embedding model to extract the fashion experts' knowledge regarding brands and incorporate it in our recommender systems. To train our embedding models, we used a curated dataset composed by brands and products descriptions written by Farfetch fashion experts, as well as highly referenced opinion articles about the brands and designers present in our catalogue.

We conducted an offline experiment to compare the embeddings generated by three well known state-of-the-art algorithms — Word2Vec, FastText and Glove [1, 10, 11]. We found that the most suitable approach to present to our customers in an online setting were the embeddings mapped by the Word2Vec model. Then, we conducted an A/B test to expose the users to the new models. We experimented with two settings: 1) recommend products from similar brands at product listing pages (PLP) following a content-based approach; 2) Boosting the current recommender systems at product detail pages (PDP) with contextual information regarding brand affinity. The results showed that this approach has greatly improved the engagement of the users with our recommendation system, leading to uplifts of up to 10% in our engagement metrics.

The main contributions of this work are 1) enlightening the importance of brand affinity in the success of luxury fashion recommendations; 2) presenting that mapping brands as word embeddings learned from fashion experts texts mirrors the fashion affinity between brands; 3) boosting recommendations with extra signals is a fast and effective way to understand the relevance of a feature in recommender systems.

2 RELATED WORKS

Embedding methods have been extensively used with great success in various Natural Language Processing (NLP) tasks [1, 10, 11]. Mikolov et al [10] proposed Word2Vec, a pair of unsupervised algorithms (Skip-Gram and CBOW) able to learn a representation for a word in a dense vector space, in which vectors representing words semantically similar are closer to each other, whereas vectors for words with semantic differences are projected further away. Joulin et al. [1] have revisited the Skip-Gram method from Word2Vec and presented FastText. This algorithm generated the semantic representations not only of the words in the vocabulary, but also of the letters and symbols composing the words in that Corpora. Finally, Pennington et al [11] proposed a new approach to learn word embeddings called GloVe (Global Vectors for Word Representation). This method is inspired by the Matrix Factorization of co-occurrence of the words in sentences. The authors claim that this approach mirrors the global semantic meaning of a word better than Word2Vec due to the word-word co-occurrence score across all Corpora.

Regarding the specific context of brand representations modelling, Yang and Cho [14] proposed the Brand2Vec approach. The work is an application of the Paragraph2Vec (or Doc2vec) algorithm proposed by Le and Mikolov [7] where the paragraph ids are the brand ids and the texts are the reviews posed by customers from a marketplace. Although Paragraph2Vec work refers the consistent superiority of PV-DM (Paragraph Vectors - Distributed Memory) over PV-CBOW (Paragraph Vectors - Continuous Bag of Words), the authors of Brand2Vec chose the PV-CBOW approach. We are not considering the paragraph modelling for this work mainly due to the interchangeability between brand representations in the texts. For example, the text written by an expert describing a particular brand or designer can include the relationships to other brands. Hence, if the brands are mapped correctly to a single token, a Word2Vec model should be sufficient to map the brand id embedding. Moreover, PV-CBOW and PV-DM are considerably more expensive computationally when comparing to Word2Vec approaches due to the need for embedding computation for the paragraph's representations.

Other automatic feature extraction techniques have been explored in a myriad of machine learning domains. In particular, on fashion related problems, Marcelino et al [9] proposed the use of a sequence model based on a Long-Short-Term-Memory (LSTM) neural network to extract features to power a semantic search engine. Unlike [9], we focused on extracting the fashion concept of what is a brand and not on parsing a general query. On a computer vision setting, one can use a Convolutional Neural Network to extract product related visual features from images and use them as side information to hybrid models to power an automated outfit generation [4, 8], or even, a more straightforward task of retrieving the nearest neighbors over the embedding space as a pure content-based filtering recommender [3]. Finally, those visual features could be employed joining with our embeddings to further improve our recommendation engines.

This work's objective is to map the expert domain knowledge from our fashion experts regarding the understating of brands. Given the high availability of textual data, we found that the approach should be focused on NLP techniques. Therefore, we selected three of the more renown methods to explore the brand embeddings learning — Word2Vec, FastText and Glove [1, 10, 11].

3 METHODOLOGY

3.1 Data collection

The main goal of this work is to create an embedding representation of the brands present in our catalogue which could mirror the fashion understanding of the experts.

For that matter, we collected five sources of text data, written in English language by fashion experts as an attempt to reflect the domain knowledge:

- Product information, accounting for more than 3M products:
  - Short description;
  - Long description;
  - Gender;
  - Category levels.
- Brand descriptions, accounting for more than 10k brands;
- Brand DNA, with top brands attributes annotated by Farfetch fashion experts on a set of 200 most popular brands:
  - Art director;
  - Fashion position of the brand;

– etc...

- Fashion Taxonomy graph, a work conducted at Farfetch leveraging fashion terms and their relationships at both product and brand level. For example, blue is an attribute to a product, but a brand with several products of the color blue will have a strong edge towards that color and can be a brand attribute. This relationship can be constructed to all the concrete and abstract fashion attributes/terms.
- Fashion articles from well renowned sources such as BoF[1], referring to brands and designers present on the Farfetch catalogue.

3.2 Data preparation

All five sources of data had the same six preprocessing steps:

(1) The text was normalized, accents and special characters were removed and the whole text converted to lowercase;
(2) All brands were mapped to a single token, if a brand name has multiple words, the space between them will be connected, eg. "Yves Saint-Laurent" maps to "yves_saint_laurent". This operation is key for the success of this implementation. If the brand name is not mapped to a single token it will be impossible to obtain a word embedding referring to a brand name. For the case of "Red Valentino", we would have two word embeddings, one for "Red" and another for "Valentino", even though in this case Red is an acronym for "Romantic Eccentric Dress" and not the colour red.
(3) The sentences were tokenized;
(4) The stopwords were removed;
(5) The resulting tokens were stemmed using Snowball method.
(6) Finally, some single tokens were joined into pairs (bigrams) in order to improve the semantic representation of frequent pairs of words.

These transformations were applied to the whole data gathered previously and a dataset of 3,806,894 sentences was obtained – the **sentence** dataset.

Two additional datasets were created from the **sentence** dataset:

- **sentence_keep**: the result from applying a filter to sentence dataset, which removes all sentences that have no term found in the Brand Names, Fashion Taxonomy, or Brand DNA — 3,263,292 sentences. The hypothesis to test with this dataset is to understand how different are the embeddings representations of a brand when using only sentences referring it.
- **sentence_keep_syn**: synonym mapping between all the words of **sentence_keep** and fashion synonyms identified by experts in Fashion Taxonomy and Brand DNA data — 3,263,292 sentences. The hypothesis to test with this data is that if we reduce sparsity, the resulting brand embeddings will have a better semantic representation.

3.3 Brand affinity modelling

Mikolov et al [10] coined the term "word embeddings" on their seminal work presenting the family of non supervised algorithms called Word2Vec. These algorithms aimed at creating probabilistic models with the objective of projecting the whole vocabulary of a Natural Language on the same multi-dimensional space. The training process can be formulated in two ways as observed in Figure 1:

[1]https://www.businessoffashion.com/

- **Skip-Gram:** given a word, the model has to be able to predict which words form its semantic context. For example, in any sentence, the model has to predict the neighbor words next to a given word.
- **Continuous Bag-Of-Words (CBOW):** given a context of N words in a sentence, the model has to predict which word is placed in that context.

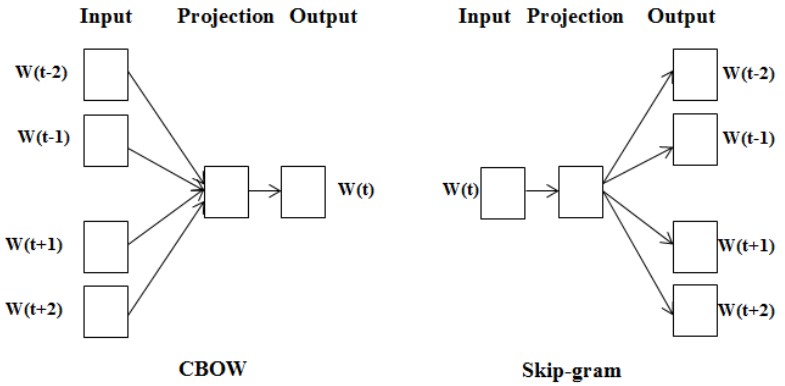

Fig. 1. Word2Vec architectures [10].

The authors observed that the semantic representation of the words using Skip-Gram approach was more accurate and at a lower computational cost [10]. Therefore, Skip-Gram architecture seems to be the most promising approach to be applied in this work for learning the word embeddings of the fashion experts data.

3.4 Boosting recommendations with brand affinity information

There are many ways to introduce side information to a recommender system. One can use it as inputs during the training process, but it requires a more complex implementation and resulting models. Another way is relying on ensemble techniques where scores from multiple models are combined into a single final score. In this work, we decided to choose the latter option. The main reason relies on experimenting the importance of a feature before investing heavily in increasing the complexity of the models currently in production.

The objective is to obtain a final score $P(i|c)$ for a given product, i from the catalogue, taking into account the context, c. Context can be any type of information regarding the current navigation intention of the user, such as gender, categories, brands, products, and so on. Hence, $P(i|c)$ can be formulated by the equation 1:

$$P(i|c) = (1 - \alpha) \cdot P'_c(i|c) + \alpha \cdot P'_b(i|c) \tag{1}$$

where P'_c is the normalized predicted score by our control recommender system; and P'_b is the normalized score given by a content-based recommender using solely the brand embeddings trained by Word2Vec in this project. Both scores are point-wise and related to given product, i, considering the context, c.

Both recommenders are separately trained and α is decided offline before an online test via click propensity optimization using logged data. α is a tunable parameter which represents the strength of the new information to the final recommendations.

4 OFFLINE EXPERIMENT

The experiments conducted in this work covered both offline and online settings to fully assess the impact of the different approaches within the recommender system's main objective of escalating our fashion experts' knowledge to our users.

4.1 Offline Setup and model selection

In this section, we cover the offline settings, where three algorithms were tested to learn the word embeddings on the aforementioned Corpora — Word2Vec, GloVe, and FastText.

For each algorithm, we tested two different sizes for the embeddings, 120 elements and 300 elements. Considering the three datasets described above (**sentence**, **sentence_keep** and **sentence _keep_syn**), 18 data-algorithm combinations were tested to understand which approach should be considered to test in a live randomized field trial, in the context of an A/B testing framework.

The offline evaluation of semantic learning is not a straightforward task due to its inherent subjectivity. We decided to follow three different evaluation approaches, one qualitative and two quantitative analysis.

First, we generated pure content-based recommendations using the embeddings obtained by the NLP algorithms by selecting the top 5 nearest neighbors for each brand present in the catalogue. Then, we used two types of pure collaborative recommenders to compare the results with. Note that these recommenders were used solely for the offline experiment.

One of the recommenders was built using navigation data to map the brand-brand relationships from user-product interactions. We built four different recommenders, one for each different source: clicks, add to wishlist, add to bag and orders. We then aggregated the results obtained by each recommender by summing the similarity scores.

The other type of recommender was based on outfit data curated by Farfetch team of stylists. From a pool of 300k outfits, we built a bipartite graph between the brands and the outfit ids to map the co-occurrence of brands in the outfit data which contains expert domain knowledge solely.

For each variation of the NLP algorithms, we computed the offline metrics such as Precision@5, Recall@5 and nDCG@5 between the auxiliary models (navigation and outfits) and the nearest neighbors based on the learned embeddings. For the final model selection, we used the Borda optimal ranking method to aggregate the different sources of results and select a single winner [2].

For subjective analysis, we used the common t-SNE projection of the embeddings to inspect the brands and their neighbors to understand if they make sense regarding the fashion context of the Corpora. This approach is standard practice and can be seen in many works where item embeddings are created [13].

4.2 Results and discussion of offline evaluation

We conducted an offline evaluation of the set of data-algorithm combinations to understand which would be the best approach to implement in a live setting and present it to our users.

Table 1 presents the results for the top 10 best combinations of the models comparing to the navigation based models referred above.

As we can see, the overall results for Precision, Recall and nDCG are very low when comparing the top-5 recommended brands by the fashion experts embeddings to the navigation-based models. Since the latter focus only on the collaborative

Table 1. Top 10 offline results comparing to navigation data.

Algo.	Size	Dataset	Prec. @5	Rec. @5	nDCG @5	Borda
Word2Vec	120	sentence	**0.0069**	0.0070	0.0238	9
Word2Vec	300	sentence	0.0067	0.0069	0.0229	8
Word2Vec	120	sentence_keep_syn	0.0067	0.0069	0.0232	7
Word2Vec	300	sentence_keep	0.0066	0.0068	0.0228	6
Word2Vec	120	sentence_keep	0.0065	0.0066	0.0232	5
FastText	300	sentence	0.0061	0.0064	0.0212	4
FastText	120	sentence	0.0061	0.0063	0.0211	3
FastText	120	sentence_keep_syn	0.0056	0.0057	0.0190	2
FastText	120	sentence_keep	0.0052	0.0054	0.0180	1
GloVe	120	sentence	0.0038	0.0038	0.0135	0

relationships between users and brands, it seems fair to assume that user interactions derive substantially different results than the recommendations obtained by the fashion experts information.

Nevertheless, the Skip-Gram Word2Vec seems to outperform the competitors regarding all metrics. Regarding the embedding size, it seems that 120 elements tend to outperform larger embedding vectors. At last, the dataset providing better metrics is **sentence** which had no extra steps of preprocessing.

Table 2 presents the results for the top 10 best combinations of the models comparing to the recommendations based on outfit data. Overall, the metrics of Table 2 are higher than those presented in the results of the navigation data (Table 1).

Table 2. Top 10 Offline results comparing to outfits data.

Algo.	Size	Dataset	Prec. @5	Rec. @5	nDCG @5	Borda
Word2Vec	120	sentence_keep_syn	**0.0101**	0.0125	0.0339	9
Word2Vec	120	sentence	0.0094	0.0117	0.0317	8
Word2Vec	120	sentence_keep	0.0090	0.0110	0.0307	7
Word2Vec	300	sentence_keep	0.0087	0.0108	0.0307	6
FastText	120	sentence_keep_syn	0.0086	0.0104	0.0285	5
Word2Vec	300	sentence	0.0086	0.0107	0.0304	4
FastText	120	sentence	0.0085	0.0107	0.0290	3
FastText	300	sentence	0.0083	0.0107	0.0279	2
FastText	120	sentence_keep	0.0079	0.0096	0.0270	1
GloVe	120	sentence	0.0046	0.0054	0.0154	0

One of the reasons for this to occur might reside in the argument that the brand embeddings proposed in this paper contain expert domain knowledge and the recommendation based on the outfit model too. Hence, it's expected that the neighbors found in both settings are more alike. Similarly to the navigation data results, the Skip-Gram Word2Vec with an embedding size of 120 seems to outperform the competitors regarding all metrics. Regarding the dataset preprocessing, **sentence_keep_syn** generates embeddings closer to the results provided by the outfit data.

Regarding the subjective analysis of the quality of the embeddings, we had projected a 2D t-SNE so we could present in this paper for reference. We plotted generic fashion terms and brands in the same space to understand if the terms

and the brands would make sense from a fashion point of view (with the help of fashion experts). For the interest of clarity, we're presenting only the embeddings projections of Word2Vec with 120 components, using the **sentence** dataset.

Figure 2 shows the brands which are neighbors to the term "cartoon". As we can observe, the nearest brands are mostly related to kids' clothing, such as Monnalisa.

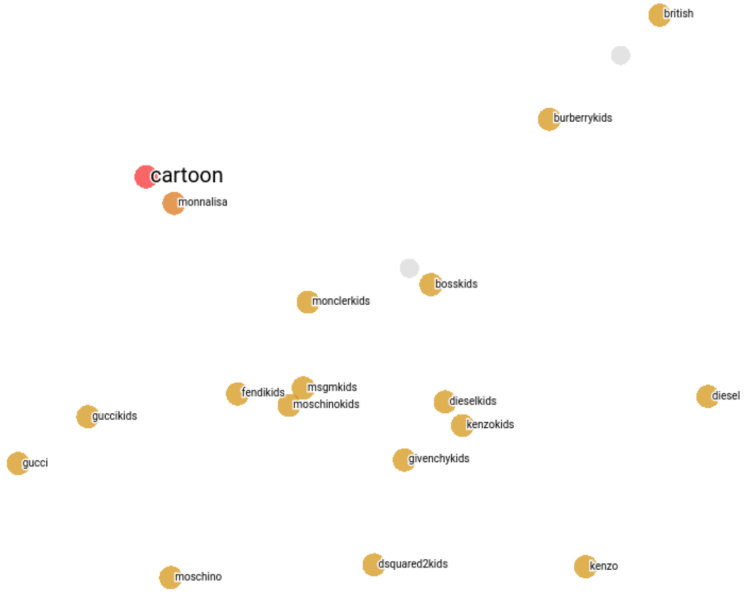

Fig. 2. t-SNE 2D projection of word embeddings emphasizing the term "cartoon".

Another example presented on Figure 3 is emphasizing the term "gothic". It's clear that the closest brand to this term is Alexander McQueen. This designer is widely known by the usage of skulls' representations in his designs.

This sort of qualitative analysis is very useful to have a general overview of how much sense the semantic representation of the words make. Overall, the embedding representations make sense. Moreover, our in-house fashion experts tended to agree on the neighbors found for a set of the brands, but, unfortunately, we don't have sufficient survey data to backup their votes and present in this paper.

The decision making process to define which model should we invest in an online experiment has considered the Borda count ranking method. When summing the Borda scores for each variation and offline experiment, we obtain the results presented in the following Table 3.

The final *Borda count* score selects as the best candidate the solution of Skip-Gram Word2Vec with an embedding of 120 elements, using the **sentence** dataset. Both first and second candidates of the final rank seem reasonable for experimenting in a live setting. Nevertheless, to support the choice of the first alternative, we present two additional arguments:

(1) We can observe a small difference regarding the offline metrics, but it does not justify the extra complexity of the ETL for processing **sentence_keep_syn**.

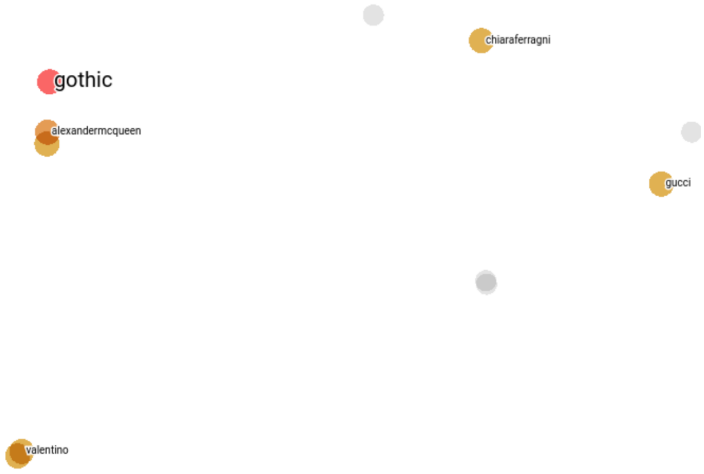

Fig. 3. t-SNE 2D projection of word embeddings emphasizing the term "gothic".

Table 3. Model selection via Borda count method.

Algo.	Size	Dataset	Borda nav.	Borda outfit	Borda final
Word2Vec	120	sentence	9	8	17
Word2Vec	120	sentence_keep_syn	7	9	16
Word2Vec	120	sentence_keep	5	7	12
Word2Vec	300	sentence_keep	6	6	12
Word2Vec	300	sentence	8	4	12

(2) The productization costs are considerably lower when annotated data, like synonyms, is not necessary.

In conclusion, we productized the following approach:

- **Algorithm:** Skip-gram Word2Vec with an embedding of 120 elements;
- **ETL:** the process to generate the **sentence** dataset, performing the transformations referred in the *Data Preparation* (Section 3.2).
- **Content-Based recommender:** using as features the embeddings obtained by the training of the Word2Vec algorithm.

5 ONLINE EXPERIMENT

A proper assessment of the impact of different approaches of a recommender system requires a variety of evaluation vectors, from objective to subjective aspects, considering user recommendation interfaces and last but not least, the ultimate intention of the user that can be affected by external factors [6]. For a thorough evaluation of the chosen final algorithm and its impact – Skip-Gram Word2Vec with an embedding of 120 elements – the online experiment was carried out in the form of four randomized field trials in a live environment. The A/B framework chosen focused on all

the users reaching the **farfetch.com** portal which were then assigned randomly (probability of 0.5) to the control or alternative groups of each of the experiments.

5.1 Online Setup

The online experiment was composed of four independent streams, as to allow a fair estimation of the algorithms' fitness to fulfill the myriad of touch-points, channels and user interaction points within its journey.

First, the resulting algorithm was interpreted as a content-based recommender on a specific set of listing pages, recommending related brands to the very specific use case of brand listings with very few items. This would be our most aggressive setting as the user's expectations were already frustrated and a successful algorithm would reconvert the user back into continuing the navigation. The null hypothesis, H_0, for this use case was then "the users are not prone to explore similar brands once their expectations have been thwarted".

Next, the algorithm was tested as a boost applied to the current product recommender systems in the following scenarios: two different types of product detail pages with the same null hypothesis, H_0, where "users are equally engaging with the control group recommendations and with recommendations that are enriched with brand affinity data".

Lastly, an edge case for our recommendation system was tested also considering the brand embeddings as a boost for the control recommender system in the form of a operational email. In this case, the user had already purchased and the goal would be to establish the fashion authority by suggesting products from related brands. The H_0 for this use case states that "the users are not susceptible to brand similarity after the purchase".

All the impressions and interactions with the recommendations carousel (as depicted in Figure 4) are recorded and a comparison of the predetermined engagement metrics[2] dictated, blindly, which alternatives could be *productized*. However, the outcomes of the four streams of online testing reflected a strong engagement gain from the users to this new source of information, across the board.

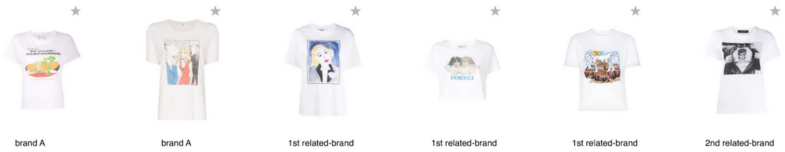

Fig. 4. Example of the recommendations carousel.

5.2 Results and discussion of online evaluation

As mentioned previously, the online experiments were executed from, essentially, two perspectives on the same brand embeddings: content-based recommendations and brand affinity boost applied on product recommenders. In table 4, a summary is presented for each of the test settings, which contributed to a full impact analysis on all aspects of the recommender system.

5.2.1 Content-based recommendations. The A/B testing framework was configured so that an even split of 50-50% of random visitors would see alternative A, the control, with no recommendations of related brand and, on alternative B, products from the top two adjacent brands we recommended at the bottom of the low-stock listing page.

[2]We reserve the right to not share the metrics in detail due to legal protection.

Table 4. Summary of the online experiments conducted.

Recommendations approach	User phase space	Touch-point
Content-based	Consideration	Low-stock listing pages
Brand Affinity Boost	Consideration	Product detail page with stock
Brand Affinity Boost	Consideration	Product detail page without stock
Brand Affinity Boost	Post-purchase	Operational Email

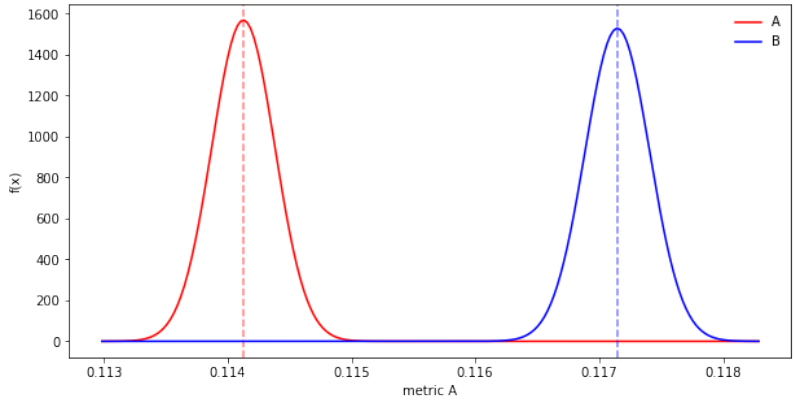

Fig. 5. A/B test results of content-based recommendations.

Figure 5 shows the PDF of the binomial derived from the logged data (impressions and conversions) of the A/B testing outcome on the low-stock PLP page. The engagement metric presented is not a click-based metric of recommendations since the control group has no recommendations to be clicked. Hence, the null hypothesis can be rejected with a p-value of 1.15×10^{-16}. The distribution of the differences between alternatives expects an engagement uplift between 1.8% and 3%, considering the confidence interval of 95%.

These results proved to be very robust in making use of, and enhancing the, subjective relationships between brands on the luxury fashion world within a context of a user that actively looked for a specific brand and was dissatisfied. Such results indicate that fashion-savvy users recognized the validity of the affinity between brands as given by the chosen algorithm to test.

5.2.2 Brand affinity boost recommendations. Using the brand embeddings as a brand affinity boost was implemented in three experiments from two distinct user phases: consideration (two experiments) and post-purchase (one experiment).

From the consideration phase, the two tests were quite similar even though they represented opposite user experiences, mainly at a layout and design level, given the two product detail pages are quite different as they represent distinct states of the product. On both, however, the same overall testing strategy was used: alternative A, the control, represented the current recommendation strategy without the boost from this new brand affinity information. On alternative B, the brand affinity boost was applied to current recommendation strategy, which was the same base strategy as alternative A.

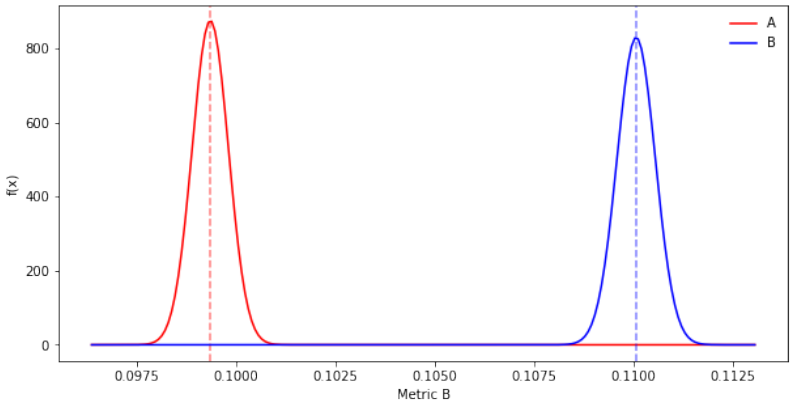

Fig. 6. A/B test results of brand affinity boosted recommendations.

Figure 6 shows the PDF of the binomial distribution derived from the logged data (impressions and conversions) of the A/B testing outcome on the PDP page. The metric presented in this A/B test was more quality of recommendations oriented than "Metric A", since the control group was also showing recommendations with the exact same recommendations carousel layout. The null hypothesis can be rejected with a p-value of 0.9×10^{-60}. The distribution of the differences between alternatives expects an engagement uplift between 7.4% and 10.6% for the product detail page without stock experiment, and for the product detail page with stock experiment the same engagement metric expected uplift was between 5.1% and 6.1%, both considering the confidence interval of 95%.

Finally, the post-purchase phase consisted of another experiment executed via email. This test has non-standard configurations and the results are not, in nature, as detailed as the traffic split of a website or device in a live environment. The experiment configuration consisted in an alternative A with a version of an algorithm that showcased only products of the same brand as an input product, whereas alternative B applied the brand affinity boost to another algorithm that did not promote same brand products. In fact, the test here allowed for a direct comparison between same brand vs related brands impact. The engagement metric revealed that the algorithm that used the brand affinity boost, alternative B, was able to outperform alternative A by approximately 15%.

The three brand affinity boost approaches tackled different combinations of user experience and user's expectation. At a product detail page level, the attention of the user is lower as it is in full exploration mode, whereas at the post-purchase phase the expectation was already fulfilled and therefore the intention to engage again is at its lowest. However, in all the three settings, the results showed the user was actively interested in the products that were of a related brand – to note this behaviour towards brands' importance has not been observed in other product attributes of previous experiments. Brand affinity as implemented in this work, proved to be successful on all randomized tests carried out to date.

6 CONCLUSION AND FUTURE WORK

In this paper, we presented an effective way of extracting and using brand embeddings, using it as side information to complement more complex recommender systems with the fashion authority expected in the luxury fashion context. The online results have shown a great acceptance from the users exposed to this information. In all the A/B tests

performed, the alternatives using the brand affinity information always won against control. The main takeaway was the understanding of brand affinity in the improvement of fashion recommender systems, in the particular context of luxury fashion.

The offline results helped to decide which approach should we choose to take to an online test. However, these decisions are often counterfactual and we cannot derive how well the other approaches would perform in a straightforward manner. Even more evident, it's when the new feature being implemented forces the re-rank of a base recommender updated regularly which drastically changes the outcome. We find it hard to foresee the outcome of an online test when the new recommender is considerably different than the control. As next steps, we plan to conduct an offline counterfactual evaluation to understand if other NLP approaches would have performed better and run the necessary online experiments to solidify our understanding. Moreover, we have A/B tests ready to start with different variations of the NLP models against the models derived for offline evaluation to understand the relationships between offline and online metrics.

As other future work, we plan as well to incorporate the brand information extracted by these embeddings in our hybrid recommender systems and other recommendation tasks such as outfits generation. We want to conduct a thorough study regarding ensemble optimization and ways to incorporate different sources of information to power recommendations in a straightforward and robust way without exploding in complexity. Finally, we plan to improve the personalization of brand affinity by considering more user navigation signals to improve the context mapping.

REFERENCES

[1] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching Word Vectors with Subword Information. http://arxiv.org/abs/1607.04606 cite arxiv:1607.04606Comment: Accepted to TACL. The two first authors contributed equally.

[2] Peter Emerson. 2013. The Original Borda Count and Partial Voting. *Social Choice and Welfare* 40 (02 2013). https://doi.org/10.1007/s00355-011-0603-9

[3] João Gomes. 2017. Boosting Recommender Systems with Deep Learning. In *Proceedings of the Eleventh ACM Conference on Recommender Systems (RecSys '17)*. ACM, New York, NY, USA, 344–344. https://doi.org/10.1145/3109859.3109926

[4] Xintong Han, Zuxuan Wu, Yu-Gang Jiang, and Larry S. Davis. 2017. Learning Fashion Compatibility with Bidirectional LSTMs. *CoRR* abs/1707.05691 (2017). arXiv:1707.05691 http://arxiv.org/abs/1707.05691

[5] Shatha Jaradat, Nima Dokoohaki, and Mihhail Matskin. 2020. Outfit2Vec: Incorporating Clothing Hierarchical MetaData into Outfits' Recommendation. In *To Appear in Special Issue (Fashion Recommender Systems) in LNSN Springer.*

[6] "B.P. Knijnenburg and M.C. Willemsen". "2015". *"Evaluating recommender systems with user experiments"* ("2nd" ed.). "Springer", "Germany", "309–352". https://doi.org/"10.1007/978-1-4899-7637-6_9"

[7] Quoc Le and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents *(Proceedings of Machine Learning Research)*, Vol. 32. PMLR, Bejing, China, 1188–1196. http://proceedings.mlr.press/v32/le14.html

[8] Yuncheng Li, Liangliang Cao, Jiang Zhu, and Jiebo Luo. 2016. Mining Fashion Outfit Composition Using An End-to-End Deep Learning Approach on Set Data. *CoRR* abs/1608.03016 (2016). arXiv:1608.03016 http://arxiv.org/abs/1608.03016

[9] José Marcelino, João Faria, Luís Baía, and Ricardo Gamelas Sousa. 2018. A Hierarchical Deep Learning Natural Language Parser for Fashion. *CoRR* abs/1806.09511 (2018). http://dblp.uni-trier.de/db/journals/corr/corr1806.html#abs-1806-09511

[10] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *CoRR* abs/1301.3781 (2013). http://dblp.uni-trier.de/db/journals/corr/corr1301.html#abs-1301-3781

[11] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*. 1532–1543. http://www.aclweb.org/anthology/D14-1162

[12] BoF team. 2017. *The 900 million dollar old Celine opportunity.* https://www.businessoffashion.com/articles/professional/the-900-million-old-celine-opportunity

[13] Laurens van der Maaten and Geoffrey E. Hinton. 2008. Visualizing Data using t-SNE.

[14] Hoseong Yang and Sungzoon Cho. 2015. Understanding Brands with Visualization and Keywords from eWOM using Distributed Representation. (2015). http://dm.snu.ac.kr/static/docs/TR/SNUDM-TR-2015-07.pdf