A Two-Phase Approach Using LDA for Effective Domain-Specific Tweets Conveying Sentiments

Pradnya Bhagat and Jyoti D. Pawar

Abstract Twitter is a free social networking platform where people can post and interact with short messages known as "Tweets". The freedom of being able to reach out to the world in a fraction of seconds has made Twitter an effective medium for the general public to express their opinion on a global scale. Since Tweets have the potential to make a global impact, companies too have started using the service to reach out to their customers. Moreover, in spite of this service being immensely effective, it is found challenging by many users to express their views through a Tweet due to the restriction imposed of minimum 280 characters. The proposed work is aimed at helping people compose better quality Tweets belonging to a specific domain in the restricted character limit. The system is designed to mine important features/topics about a domain using Latent Dirichlet Allocation (LDA) algorithm and to compute the polarity of the sentiment words associated with them with respect to the domain using a two-phase approach on an Amazon review corpus. The discovered topics/features and sentiments are recommended as suggestions to Twitter users while composing new Tweets. The paper describes and presents initial results of the system on cell phones and related accessories domain.

Keywords Social networking · Twitter · E-commerce · Product review Tweets · Recommendations · Topics · Sentiment words · Latent Dirichlet Allocation

1 Introduction

Social media has started playing a pivotal role in our day-to-day life. Facebook, Twitter, Instagram, LinkedIn are examples of some of the networks developed to cater to the rising and constantly changing needs of the society [1]. Twitter [15] is

P. Bhagat (✉) · J. D. Pawar
Goa Business School, Taleigao, Goa, India
e-mail: dcst.pradanya@unigoa.ac.in

J. D. Pawar
e-mail: jdp@unigoa.ac.in

J. K. Mandal et al. (eds.), *Computational Intelligence and Machine Learning*, Advances in Intelligent Systems and Computing 1276,
https://doi.org/10.1007/978-981-15-8610-1_9

Fig. 1 Examples of Tweets about products [15]

a platform where users can interact with each other with the help of short messages knows as "Tweets". It is basically a micro-blogging platform where the message length is restricted to 280 characters. Since it is a free service with a global reach, people all over the world have started taking advantage of the same. Of late, we have seen an increase in the trend by people to review various products/services used by them on Twitter [5]. Twitter is also emerging as a platform to lodge complaints about goods or services having suffered the inconvenience. Being a broadcasting medium, the Tweeted message has the potential to instantly reach thousands of people creating global impact. Many other users reading this Tweet can base their purchase decisions on the read Tweet. This forces the companies to give immediate attention and address the issues faced by customers. Hence, Twitter is emerging as an effective medium for the public to make their problems heard and addressed. Figure 1 shows examples of some Tweets written by users addressed to the companies.

Although being such a powerful platform, we see that the use of Twitter is mostly restricted to specific sections of society. It has still not become successful in reaching to general masses. One of the reasons can be attributed to the 280 character limit for each Tweet, which acts as a great challenge to express one's opinions in brief. Especially to people whom English may be a foreign language, expressing one's views in such a constrained manner may not be always possible. The proposed system is designed with an aim to aid people; compose better quality Tweets related to products by recommending them with various product features and domain-specific sentiment words while composing Tweets. The system is unique from the fact that recommendations to help generate Tweets are sourced from amazon.com [2] which is an e-commerce website. Since there is no restriction on the length of the reviews on Amazon, it is found that Amazon reviews are more informative and descriptive than Tweets. Hence, reviews are used to generate recommendations. The discovery of topics from reviews is done using Latent Dirichlet allocation (LDA) [4] algorithm. The paper also proposes a method to find the polarity of sentiment words with respect to a particular domain since the sentiment of a word can largely depend on the context used. Since we are using the e-commerce domain to scrap reviews in order to develop

recommendations for writing on Twitter which is a social networking platform, we can say that the proposed system is also a step towards cross-domain recommender systems.

2 Literature Survey

A significant amount of literature is being continuously contributed in the fields of social networking and e-commerce to keep up with the ever-evolving areas. The research depends heavily on Natural Language Processing and Machine Learning algorithms.

Dong et al. [7] describes a tool called Reviewer's Assistant that can be added as a browser plug-in to work with e-commerce websites like Amazon to help users write better quality reviews. The tool mines important topics from already published reviews and presents them as suggestions to new writers. Blei et al. [4] presents the Latent Dirichlet Allocation (LDA) algorithm, a generative probabilistic model for a collection of discrete data such as document collections. Dong et al. [6] describes a novel unsupervised method to extract topics automatically from a set of reviews. The work uses LDA algorithm and presents a major improvement on previous methods which required manual intervention. ki Leung et al. [10] proposes a probabilistic rating inference model for mining user preferences using existing linguistic processing techniques from a set of reviews and mapping the preferences on a rating scale. The method allows semantically similar words to have different sentiment orientations and hence tries to address the limitations of existing techniques. Zhang and Liu [16] presents a method to identify product nouns that imply opinions based on statistical tests.

3 Methodology

The work addresses the problem of assisting users to write better quality reviews using a two-phase approach:

1. Recommending topics/features in which the user may wish to express his/her opinions.
2. Suggesting appropriate opinion words to convey sentiments about the stated topics/features.

3.1 Recommending Product Features

Generally, many users are unaware of the technical terminology related to products. As a result, they are unable to use the correct terminology to explain the problem technically. Moreover, on a platform like Twitter, using elaborate sentences to explain one's problem is not feasible due to the character limit imposed. The proposed method uses LDA algorithm to address the issue. The approach adopted is as follows:

Reviews are broken down into sentences since most of the times a single sentence describes a single topic. The extracted sentences are Parts-of-Speech (POS) Tagged [13, 14] to identify the various parts of speech in the review. Since, most of the times, nouns are the parts of speech that convey the topic/features of the products, only nouns need to be retained from the entire review text. The challenge lies in identifying feature nouns from non-feature nouns.

In general, it is seen that the feature nouns occur in close proximity to sentiment words since the users are interested in expressing their opinions about the same. This is not the case with non-feature nouns. For example:

My friend advised me to buy this awesome mobile because it has this stunning look and attractive features.

POS Tagging of the above sentence would give us:

My_PRP friend_NN advised_VBD me_PRP to_TO buy_VB this_DT awesome_JJ phone_NN because_IN it_PRP has_VBZ this_DT stunning_JJ look_NN and_CC attractive_JJ features_NNS.

The nouns occurring in the above text are *friend*, *phone*, *look* and *features*. Out of these, the nouns we would be interested are *phone*, *look* and *features* since they belong to features/topics about cell phones. Further, as can be seen, nouns *phone*, *look* and *features* have some adjective associated with them since the user wants to express his opinions about the features, but noun *friend* does not have any adjective associated with it. This observation is utilized to differentiate feature nouns from non-feature nouns.

Next, the sentence position of the feature nouns is retained and the pre-processed file is given to the LDA algorithm. LDA is a statistical algorithm used to automatically identify topics across documents [4]. The algorithm follows the bag-of-words approach and considers documents as a set of topics that are made up of words with certain probabilities. As per the working of the algorithm, the words forming similar topics get grouped together. Whenever a person starts to compose a Tweet on any topic, many other words related to the same topic get displayed to the user. This can act as a valuable aid to the users to get correct technical words to express information effectively in the specified character limit.

3.2 Suggesting Appropriate Opinion Words

The second important part of the proposed framework deals with suggesting sentiment words or opinions about the topics to be addressed. Most of the works in the literature focus on selecting the adjectives based on POS Tagging and comparing them against a sentiment lexicon to find the polarity (positive/negative/neutral) of the opinions.

Although in the majority of the cases the approach works fine, it fails to identify the sentiment words whose polarity may be dependent on the context of usage. A classic example in sentiment analysis is the word *unpredictable* which has a clearly negative polarity if used in car domain: *The steering wheel of the car is unpredictable.* But is found to have a positive polarity in the movie domain: *The ending of the movie is unpredictable.* As a result, it can be stated that the context of using the word plays a major role in determining its polarity. The method followed in the paper is distinct in a way that it does not just follow the polarity of the words as given in the sentiment lexicon; instead, it computes it from the occurrence frequency of the sentiment word in the review dataset. The steps followed are as follows:

All adjectives are extracted from the POS tagged data. It need not be the case that all adjective words carry some sentiment meaning. For example, if we have the phrases:

front flash and *good flash*

in our reviews, POS tagging will tag both as:

front_JJ flash_NN and *good_JJ flash_NN*

As can be seen, both examples get reduced to Adjective(JJ)-Noun(NN) phrases. The phrase *good flash* carries some sentiment (positive) associated with it, whereas *front flash* does not. Hence, to identify whether an adjective is a sentiment word or not, we first make use of a sentiment lexicon [9] to find the presence of a specific adjective in the sentiment lexicon. If the adjective is present in the sentiment lexicon, we consider that word to be a sentiment word and vice versa. The sentiment lexicon is used only to identify if an adjective is a sentiment word or not. It is not used to actually identify the polarity of the sentiments.

The review dataset has a numeric rating associated with every review given by the reviewers in addition to the review text. The rating is in the form of stars and it ranges from 1 star to 5 stars. We group the reviews according to the number of stars associated with the review. To find the polarity of the sentiment words, we take the sentiment words extracted using the sentiment lexicon and find their occurrence across all five groups of reviews. If any sentiment word occurs the majority of the times in 4/5 star reviews, we consider that word to be extremely positive. If any

sentiment word occurs the majority of the times in 1/2 star reviews, we consider that word to be extremely negative. If any word occurs an equal number of times in both, positive and negative reviews, we consider that word to be neutral in polarity. Hence, using this method we get the context-based sentiment polarity of the words.

4 Implementation and Experimental Results

The proposed work is implemented using Python programming language [12]. The preprocessing operations and POS tagging is carried out using the Natural Language Processing Toolkit (NLTK) [3]. Gensim topic modelling library [11] is used for LDA implementation.

The dataset for the study presented is sourced from Amazon [8]. It is a collection of Amazon reviews on the topic *Cell Phones and its related Accessories*. The first part of the work consists of the use of LDA to extract groups of words forming similar topics so that whenever the user starts Tweeting about a topic, all the related words get displayed to him/her to help him make his Tweet more informative. The optimal number of topics for the stated dataset was found to be 5 with five words in each topic. The algorithm was run for 10 passes. As we increased the number of topics and words, it is observed that the topic clusters lost their coherence; hence, we restrict the number of topics and number of words to 5.

The topics along with the words obtained are displayed in Table 1. Topic 1 corresponds to battery or charging. Topic 2 describes the case or the protection of the phone. The third topic can be estimated to vaguely describe the protector of the screen. The fourth topic is about cables or ports of mobile phones. The fifth topic, as can be seen, describes the overall quality of the phone. Hence, whenever a user types any one word from the topic, we can deliver him suggestions with other words in the same topic to help him make his tweet more informative.

The next part of the work deals with identifying sentiment words with the correct sentiment orientations in the concerned domain. Table 2 displays the normalized occurrence frequency of sentiment words across various review categories. As can be seen words like *poor, less, cheap* have higher occurrence frequency in 1/2 star

Table 1 Topics discovered using LDA

Topic 1	Battery	Device	Phone	Charger	Charge
Topic 2	Case	Protect	Color	Plastic	Phone
Topic 3	Protector	Screen	Thing	One	Part
Topic 4	Cable	Port	Car	Cord	Tip
Topic 5	Quality	Product	Time	Price	Sound

Table 2 Normalized occurrence frequency of sentiment words across review categories

1 Star	2 Star	3 Star	4 Star	5 Star
Less 1	Right 1	Bad 0.57269	Last 0.622323	Excellent 1
Poor 1	Second 0.452949	Cheap 0.355925	Clear 0.592842	Happy 1
Second 0.547051	Bad 0.297234	Clear 0.277108	Good 0.422779	Perfect 0.800259
Bad 0.42731	Cheap 0.267953	Full 0.235836	Little 0.301183	Good 0.703959
Cheap 0.376122	Last 0.170843	Good 0.172343	Nice 0.281524	Great 0.695143
Last 0.206834	Clear 0.13005	Hard 0.15393	Hard 0.27927	Easy 0.648894
First 0.095487	Better 0.079944	Better 0.153892	Long 0.266587	Best 0.640565
Better 0.064655	First 0.077419	First 0.114706	Easy 0.260472	New 0.637654
New 0.0644	Hard 0.077419	Little 0.107598	Best 0.248526	Long 0.632026
Good 0.053394	Good 0.070304	Nice 0.107075	Better 0.248526	Full 0.614961
Full 0.052943	New 0.053343	Long 0.101387	Full 0.235836	Nice 0.55073
Hard 0.052006	Nice 0.040275	Easy 0.090634	First 0.228803	Little 0.52144
Little 0.030312	Little 0.039468	New 0.085095	New 0.203437	First 0.508044
Nice 0.020395	Easy 0.029232	Best 0.069352	Perfect 0.199741	Better 0.452983
Great 0.018186	Great 0.027475	Great 0.060453	Great 0.198784	Hard 0.437375

reviews, hence, it can be inferred that these words are negative in polarity. On the other hand, words like *excellent, happy, great* have higher occurrence frequency in 4/5 star reviews; conforming their positive polarity.

5 Conclusion and Future Work

The work presented a method to assist users in the task of Tweeting informative messages in the required character length to make Twitter easier for the general public. The LDA algorithm used for topic modelling groups related words belonging

to a common topic together from the dataset which can be displayed to the Twitter users as help while he/she is typing the Tweet. The problem of different words having different sentiment orientations in different domains is also taken care of since the method doesn't totally depend on sentiment lexicon to find sentiment orientations. Instead, the method computes the sentiment orientations of words based on the occurrence frequency in various classes of reviews. The future work consists of the development of the entire system and a live user trial to find the effectiveness of the system in the real world.

Acknowledgements This publication is an outcome of the research work supported by Visvesvaraya PhD Scheme, MeitY, Govt. of India (MEITY-PHD-2002).

References

1. Alhabash, S., Ma, M.: A tale of four platforms: Motivations and uses of facebook, twitter, instagram, and snapchat among college students? SAGE Publications (2017)
2. Amazon: https://www.amazon.com/
3. Bird, S., Loper, E.: Nltk: the natural language toolkit. In: Proceedings of the ACL 2004 on Interactive Poster and Demonstration Sessions (2004)
4. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**, 1022–1993 (2003)
5. Curran, K., O'Hara, K., O'Brien, S.: The role of twitter in the world of business. Int. J. Bus. Data Commun. Netw. (2011)
6. Dong, R., Schaal, M., Ad Kevin McCarthy, M.P.O., Smyth, B.: Unsupervised topic extraction for the reviewer's assistant. In: International Conference on Innovative Techniques and Applications of Artificial Intelligence (2012)
7. Dong, R., Schaal, M., O'Mahony, M.P., Smyth, B.: Topic extraction from online reviews for classification and recommendation. In: Twenty-Third International Joint Conference on Artificial Intelligence (2013)
8. He, R., McAuley, J.: Ups and downs: modeling the visual evolution of fashion trends with oneclass collaborative filtering. WWW (2016)
9. Hu, M., Liu, B.: Mining and summarizing customer reviews. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2004)
10. ki Leung, C.W., Fai Chan, S.C., Lai Chung, F., Ngai, G.: A probabilistic rating inference framework for mining user preferences from reviews. World Wide Web **14**(2) (2011)
11. Řehřek, R., Sojka, P.: Gensim—statistical semantics in python. Statistical semantics; gensim; Python; LDA; SVD (2011)
12. Rossum, G.: Python reference manual (1995)
13. Taylor, A., Marcus, M., Santorini, B.: The penn treebank: An overview. In: Abeille A. (eds) Treebanks. Text, Speech and Language Technology, vol. 20. Springer, Dordrecht (2003)
14. Toutanova, K., Klein, D., Manning, C., Singer, Y.: Feature-rich part-of-speech tagging with a cyclic dependency network. In: Proceedings of HLTNAACL (2003)
15. Twitter: https://www.twitter.com/
16. Zhang, L., Liu, B.: Identifying noun product features that imply opinions. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (2011)