Social Media Chatbot System - Beekeeping Case Study

Zine Eddine Latioui, Lamine Bougueroua, Alain Moretto

▶ **To cite this version:**

Zine Eddine Latioui, Lamine Bougueroua, Alain Moretto. Social Media Chatbot System - Beekeeping Case Study. Madureira A.; Abraham A.; Gandhi N.; Varela M. Hybrid Intelligent Systems, HIS 2018, 923, Springer International Publishing, pp.302-310, 2020, Advances in Intelligent Systems and Computing (AISC), 978-3-030-14346-6. 10.1007/978-3-030-14347-3_29 . hal-04068874

HAL Id: hal-04068874
https://hal.science/hal-04068874v1

Submitted on 14 Apr 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Copyright

Social Media Chatbot System - beekeeping case study

Zine Eddine Latioui, Lamine Bougueroua, Alain Moretto

AlliansTic, Efrei Paris, Villejuif, France
{ zine.eddine.latioui, Lamine.bougueroua, Alain.moretto}@efrei.fr

Abstract. The aim of this paper is to present an innovative way for assisting beekeepers during the process of taking care of their apiary based on text mining and deep learning. To reach this goal, we propose an innovative social media *Chatbot* called *ApiSoft*. This system is able to extract relevant information by processing data from different sources like social media, web, data provided by expert and our applications embedded on the beekeepers' smartphone. Once data are collected, *ApiSoft* can sand alerts, information and pieces of advice about the state of apiaries to all subscribers according to their specific interests. We believe that this approach will not only lead to a better monitoring of production but will also allow an enhanced monitoring of the sector at regional and national level.

Keywords: Text mining, deep learning, Social media data, Chatbot.

1 Introduction

Social Networks are indisputably popular nowadays and show no sign of slowdown. According to the Kepios study [1], the number of active users of social networks increased by 13% in 2017 to reach 3.3 billion users in April 2018. For example, Facebook attracts more than 2.2 billion users a month. Penetrating ever more aspects of our daily life, they become not only a considerable threat for our privacy, but also an encompassing tool for analyzing opinions, habits, trends and some would even say – thoughts.

In the current growth of artificial intelligence, machine learning and natural language processing, driven by new technological possibilities, it is possible to automate the analysis of vast amounts of publicly published data.

Text Mining and Social Network Analysis have become a necessity for analyzing not only information but also the connections across them. The main objective is to identify the necessary information as efficiently as possible, finding the relationships between available information by applying algorithmic, statistical, and data management methods on the knowledge. The automation of sentiment detection on these social networks has gained attention for various purposes [2][3][4].

Sentiment Detection, or in its simplified form – Polarity Classification, is a tedious and complex task. Contextual changes of polarity indicating words, such as negation,

sarcasm as well as weak syntactical structures make it troublesome for both machines and humans to safely determine polarity of social media messages.

Twitter is a social network that allows the user to freely publish short messages, called Tweets via the Internet, instant messaging or SMS. These messages are limited to 140 characters (more exactly, NFC normalized codepoints[5]). With about 330 million monthly active users (as of 2018, Twitter Inc.), Twitter is a leading social network, which is known for its ease of use for mobile devices (90% of users access the social network via mobile device). Note also, the more mature age of Twitter users: the most represented age group is 35-49 years old. Twitter known by the diversity of content, as well as its comprehensive list of APIs offered to developers.

With an average of 500 million messages sent per day, the platform seems ideal for live tracking opinions on various subjects. Furthermore, the very short format messages facilitate classification since short messages rarely discuss more than one topic. However, automated interpretation is complicated by embedded links, abbreviations and misspellings. Facing these challenges is becoming increasingly important for Economic and Market Intelligence in order to successfully recognize trends and threats.

We are interested in a particular use case; it concerns the sector of beekeeping. The importance of honeybees in agriculture has gained public attention in recent years, along with wide news coverage of their decline. Growing numbers of people are becoming concerned about the plight of honeybees. According to *FranceAgrimer* [6], the average age of beekeepers in France is 42 years old.

The beekeeping industry faces many problems related to the health of bee colonies concomitant with a general decline in production. Scientific work focused on understanding these phenomena has allowed identifying several causes and establishing concrete elements to guide the decisions of beekeepers.

Due to pollution and climate change, experts confirmed that the massive destruction of marine creatures could occur very soon. Many terrestrial species are also exposed to the same fate. The death of the honeybees' population is considered the most important warning to humanity, as they play a more vital role in pollination (this is an important step in horticulture and agriculture).

In this paper, we will present an original method in order to help beekeepers take care of their apiary. Once consistent data from different sources are collected, a deep learning AI is run to give beekeepers advices trough a smart Chatbot.

This paper is structured as follows. In the first place, we discuss the state of art then we pass to describe the utilities and functionalities of our smartphone application in section 3. In section 4, we talk about text mining and data aggregation, and we end with a conclusion and some perspectives.

2 Related work

This idea is inspired by previous work in the domain of machine learning especially in deep learning where we could find a lot of interesting deep neural net applied to speech recognition needed in our application.

We can mention for example [7][8][9], however we will be using Baidu deep speech 2 [10] because they have proved its performance over two different languages (English and Mandarin). Since the approach is highly generic, it can be quickly be applied to new languages. Therefore, we suppose that it will be suitable for other languages like the French language, which is our main language.

Text mining is an active research area when it comes to web and social media data, as mentioned in [11][12], people on social media such as Twitter don't pay much attention to their spelling or grammatical construction faults, as a result preprocessing these data is an irreversible step towards a valid mining system.

In this regards we intend to do as in [13], so we will be using an n-gram model to correct the spelling and grammar faults, we replace the emoticons with their meaning and also we will remove punctuation, symbols, links, hashtags, targets, replace opinion phrases and idioms with their truth meaning, and also filter the language used in the tweet. Since we want to use machine-learning techniques to deal with this problem we first need to extract feature from the preprocessed tweets.

Different type of features are used in the literature, we note for example: using an n-gram model to link words with their negative and positive probabilities.

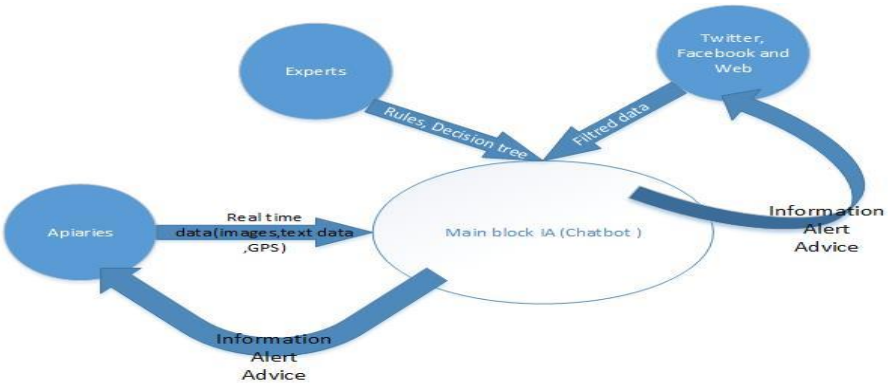
3 System architecture

The core of our system is a Chatbot capable of sending valuable information to subscribed beekeepers by collecting and extracting data from blocks and from deferent sources. So that it could use the collected data to train machine learning model capable of generating advice or alerts to users in human understandable language.

Figure 1 shows the structure of our proposed model we can see that this system is composed of different parts; it consists of four major components:

- apiary(smartphone application),
- social media and web mining,
- expertise rules,
- Chatbot.

Fig. 1. General model structure.



(i) We use our application to collect real-time information from beekeepers that uses our application to control their apiaries.

(ii) We applied text-mining approach to extract data from social media such as Twitter and filter it to keep the most relevant information.

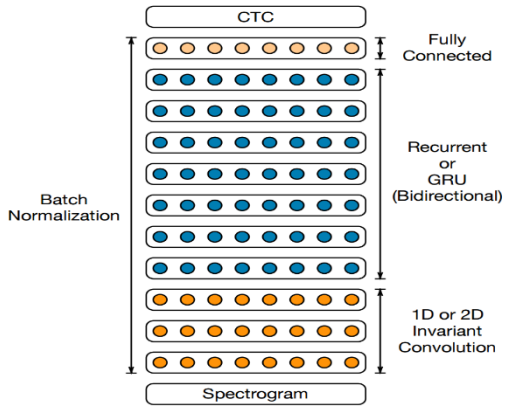
(iii) At this stage, we use expert's knowledge to form some sort of rules. We thus create an expert agent.

(iv) The Chatbot ApiSoft is the main component of our system. Its mission is to address beekeepers with eventual alerts or valuable information based on data provided from all other parts of the system.

A. Smartphone application (Apiaries)

The reason for creating this application is to help beekeepers manage their apiaries. At the same time, we can use data collected by this application to feed the Chatbot. For example, if there is a certain bee disease that has been located in a geographic area, then the Chatbot can notify all nearby beekeepers about that disease.

Fig. 2. Deep neural network used for speech recognition [10].



In this application, we have proposed a speech recognition system inspired by Baidu deep speech 2[10]. Figure 2 presents the structure of deep neural network we use. The first layers consist of convolutional ones since they have proved capable of extracting pertinent features.

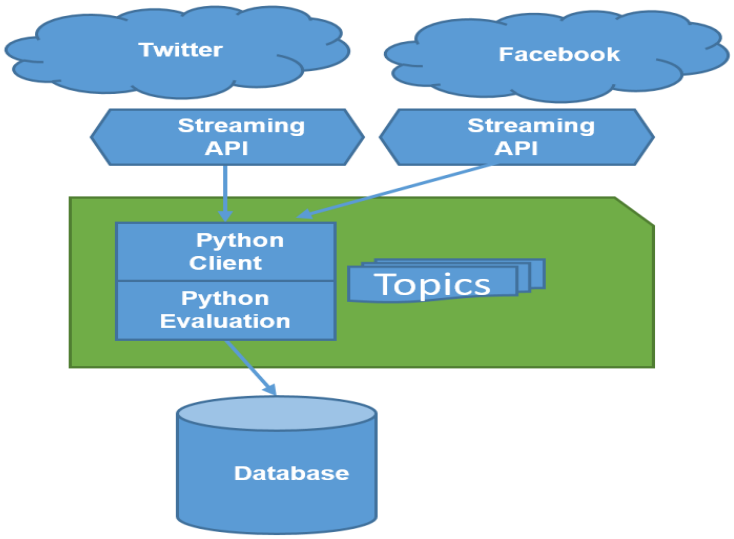
After the convolutional step, we implemented bidirectional recurrent layers using Gated Recurrent Units (GRUs). GRU units are preferred to LSTMs since they consume much less energy and are easier to train compared to LSTMs. We finish with a one fully connected layer. Furthermore, we have used Connectionist Temporal Classification (CTC) loss function [14] to train our model. For performance reasons, we chose to use WrapCTC [10]. The most challenging part in this deep learning approach was to find data for training since deep learning required more training data then traditional approaches.

With this Speech Recognition (SR) system, users are capable of wearing there bee uniforms and communicating commands and information about their beehives to this application over their voice, at the same time. They do not need to touch their phones so they are not bothered in their movements. One additional utility of this application consists in processing images data such in order to detect and count Varroa mites.

B. **Social media and web mining**

The block, described in figure 3, is responsible for collecting topic related data from the web particularly from social media like Twitter or Facebook. They represent major sources of data. Because of its data type and the number of users, we use a python client interface to connect to twitter API to search for tweets on a specific topic. For example, we can search for tweets containing words like apiary, honey, beekeeping, etc.

Fig. 3. Proposed structure for collecting and mining social media network data.



Twitter which is proved to be a great source of information [15][16] since it contains a large community and only handles short text data, which make it ideal of text mining techniques.

We perform an evaluation of the collected tweets, in which we will be using a probabilistic n-gram model to indicate the polarity of the tweet. This mechanism will allow us to extract knowledge or patterns and store it in our database so that we can use it to generate advice or alerts. Chatbot communicates all information to users.

C. Expertise rules

In this part, we take bee experts knowledge to form a sort of rules. We use these rules to build a decision tree. We uses this tree on information provided from other blocks, which can help us generate meaningful warnings and assistance.

We came to choose a decision tree since despite it does not require too much data; it is capable of handling multi output problems. Besides, the cost of implementing a tree is logarithmic which a gain in complexity.

First we begin by creating a web survey addressed to bee experts, to collect their knowledge in a format that we want, after that we pass to train the model using C5.0 which is a sophisticated data mining tool generally describe as if-else- rules [17].

D. Chatbot

This is the main component of our system. It uses all information and approaches accumulated from other blocks, to generate human comprehensive text.

To create a system capable of generating human understandable content we will be using a sequence-to-sequence neural net. We came to this chose as these models provide versatility compared to traditional machine learning methods that required a fixed output length. The generated message is then broadcasted to all concerned beekeepers so that it can help them while they maintain their apiary.

4 Results and Discussion

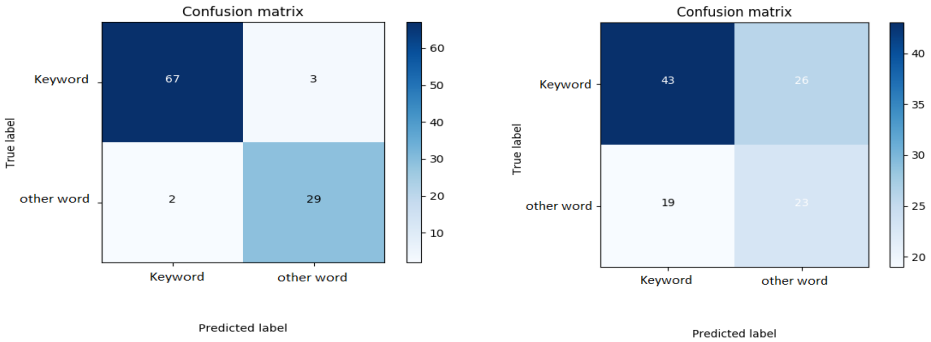
In this study, we are interested in the voice recognition part. We first, we speak about *hotword* detection where we try to detect specific word if this word is pronounced. In order to select the suited system we tested two methods *Snowboy* detection and *CMU Sphinx*.

Snowboy: is a personalized real-time deep neural network capable of identifying *hotword* from continues speech. The advantage of this library is that it uses insignificant computing power that makes it capable of working even in Raspberry Pi of the first generation.

The second toolkit is *CMU Sphinx,* which is mainly a full speech recognition system. It also offers a special mode that makes it able to detect targeted words.

We tested the performance between this two libraries by a dataset containing words like hey google, *Apisoft* etc. we have used confusion matrix (shown in figure Fig.4) as a metric of performance evaluation.

Fig. 4. Confusion matrix



(a) Snowboy (b) CMU Sphinx

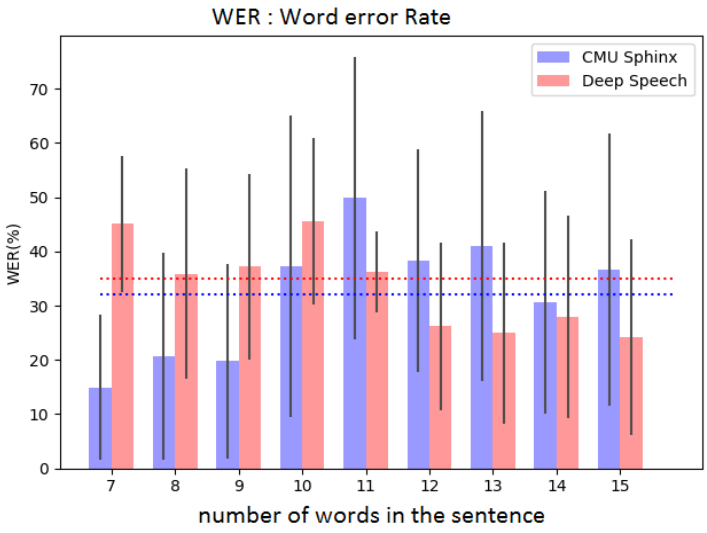
For speech recognition, we have also tested two approaches. In orders to discuss whatever to choose of using *CMU sphinx* completes speech recognition or an end-to-end deep neural network.

The first approach is based on the classical pipeline of a typical speech recognition system which counting an acoustic model. In the case of *CMU Sphinx*, they have used a Hidden Markov model capable of converting a set of audio features into a list phonemes, the second component is a dictionary to do the mapping between the phonemes the worlds, the last component is the language model it's trying to minimize vocabulary and grammatical errors.

The second approach is a deep recurrent neural network that we trained based on deep speech two architecture we have used 50 hours of training data (mainly VoxForge dataset plus books reading samples). Since we do not really have a large amount of data, we just keep the network as simple as possible with three convolutional layers followed by a bidirectional neural network.

We did a comparison of these two methods described above. the figure 5 show the WER (Word error Rate) by the length of sentences for the two systems we can see that CMU Sphinx, in general, is still a bit better than the end to end DNN approach with 32% compared to 35% for the neural network. We think that this is due to the quantity of data we have. Since the end to end, DNN approaches generally required a larger dataset to train. However, we will interpret the data problem, considering we have some collaborators that are willing to give us the necessary data to train a <10% WER Speech recognition system based on this approach.

Fig. 5. Word error Rate by the length of sentences



5 Conclusion

In this paper, we have presented a novel way capable of helping beekeepers to control their apiary by communicating to users' notifications about the state of apiaries in their respective areas and inform them if there is a phenomenon to consider.

Our method based on machine learning and text mining techniques tries to build a Chatbot that collects data from various sources such as social media, experts and real data (different types of sensors in hives) collected by smartphone application. We think this approach will continue to improve as this system continues to work and training.

At last, we expect that this system will give valuable aids in the beekeeping process. We also hope that our solution will contribute to the improvement of the quantity and quality of honey production and will participate in the prevention against the extinction of honeybees.

References

1. Kepios, "Digital in 2018, essential insights into internet, social media, mobile, and ecommerce use around the world". April 2018. [https://kepios.com/data/]
2. M. Ghiassi, J. Skinner, D. Zimbra, "Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network". Expert Systems with Applications, Volume 40, Issue 16, 15 November 2013, Pages 6266-6282.
3. X. Zhou, X. Tao, J. Yong, Z. Yang, "Sentiment analysis on tweets for social events". Proceedings of the 2013 IEEE 17th Int. Conference on Computer Supported Cooperative Work in Design. CSCWD 2013. 27-29 June 2013. pp 557–562.

4. M. Salathé, D.Q. Vu, S. Khandelwal, D.R. Hunter, “The dynamics of health behavior sentiments on a large online social network”. EPJ Data Science 2 : 4. 2013, 1–12. doi: 10.1140/epjds16.
5. B. Sriram, D. Fuhry, E. Demir, H. Ferhatosmanoglu, M. Demirbas, “Short text classification in twitter to improve information filtering”. Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval. July 19 - 23, 2010. Pages 841-842. doi: 10.1145/1835449.1835643.
6. Proteis+, “Audit economique de la filière apicole française". Final report, 2012. FranceAgriMer.
7. A. Graves, J. Navdeep, "Towards end-to-end speech recognition with recurrent neural networks". In International Conference on Machine Learning, pp. 1764-1772. 2014.
8. A. Maas, Z. Xie, D. Jurafsky, A. Ng, "Lexicon-free conversational speech recognition with neural networks". In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 345-354. 2015.
9. W. Chan, N. Jaitly, Q. Le, O Vinyals. "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition". In Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on, pp. 4960-4964. IEEE, 2016.
10. D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper et al. "Deep speech 2: End-to-end speech recognition in english and mandarin." In International Conference on Machine Learning, pp. 173-182. 2016.
11. R. Parikh, M. Movassate Sentiment analysis of user-generated twitter updates using various classification techniques. CS224N Final Report. 2009 Jun 4; 118.
12. B. Gokulakrishnan, B. Priyanthan, T. Ragavan, N. Prasath, A. Perera, Opinion mining and sentiment analysis on a twitter data stream. InAdvances in ICT for emerging regions (ICTer), 2012 International Conference on 2012 Dec 12 (pp. 182-188). IEEE.
13. V. Kharde, P. Sonawane . Sentiment analysis of twitter data: A survey of techniques. arXiv preprint arXiv:1601.06971. 2016 Jan 26.
14. A. Graves, S. Fernández, F. Gomez, J. Schmidhuber. "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks." In Proceedings of the 23rd international conference on Machine learning, pp. 369-376. ACM, 2006.
15. L. Barbosa, J Feng. "Robust sentiment detection on twitter from biased and noisy data." In Proceedings of the 23rd international conference on computational linguistics: posters, pp. 36-44. Association for Computational Linguistics, 2010.
16. A. Bifet, E. Frank. "Sentiment knowledge discovery in twitter streaming data." In International conference on discovery science, pp. 1-15. Springer, Berlin, Heidelberg, 2010.
17. S. Hou,R. Hou,X. Shi,J. Wang,C. Yuan. “Research on C5. 0 algorithm improvement and the test in lightning disaster statistics”. International Journal of Control and Automation. 2014; 7(1):181-90.