Audience Segmentation in Social Media

Verena Henrich, Alexander Lang
IBM Germany Research and Development GmbH
Böblingen, Germany
verena.henrich@de.ibm.com, alexlang@de.ibm.com

Abstract

Understanding the social media audience is becoming increasingly important for social media analysis. This paper presents an approach that detects various audience attributes, including author location, demographics, behavior and interests. It works both for a variety of social media sources and for multiple languages. The approach has been implemented within *IBM Watson Analytics for Social Media*™ and creates author profiles for more than 300 different analysis domains every day.

1 Understanding the Social Media Audience – Why Bother?

The *social media audience* shows *Who* is talking about a company's products and services in social media. This is increasingly important for various teams within an organization:

Marketing: *Does our latest social media campaign resonate more with men or with women? What are social media sites where actual users of our products congregate? What other interests do authors have that talk about our products, so we can create co-selling opportunities?*

Sales: *Which people are disgruntled with our products or services and consider churning? Can we find social media authors interested in buying our type of product, so we can engage with them?*

Product management and product research: *Which product features are important specifically for women, or parents? What aspects do actual users of our product highlight—and how does this compare to the competition's product?*

Besides commercial scenarios, social media is becoming relevant for the social and political sciences to understand opinions and attitudes towards various topics. Audience insights are key to put these opinions into the right context.

2 Audience Segmentation with IBM Watson Analytics for Social Media

IBM Watson Analytics for Social Media™ (WASM) is a cloud-based social media analysis tool for line-of-business users from public relations, marketing or product management[1]. The tool is *domain-independent*: users configure *topics* of interest to analyze, e.g., products, brands, services or politics. Based on the user's topics, WASM retrieves all relevant content from *a variety of social media sources* (Twitter, Facebook, blogs, forums, reviews, video comments and news) across *multiple languages* (currently Arabic, English, French, German, Italian, Portuguese and Spanish) and applies various natural language processing steps:

- Dynamic topic modeling to spot ambiguous user topics, and suggest variants.
- Aspect-oriented sentiment analysis.
- Detection of spam and advertisements.
- Audience segmentation, including author location, demographics, behavior, interests, and account types.

While the system demo will show all steps, this paper focuses on *audience segmentation*. Audience segmentation relies on: (i) The author's *name* and *nick name*. (ii) The author's *biography*: a short, optional self-description of the author (see Figure 1), which often includes the author's location. (iii) The author's *content*: social media posts (Tweets, blog posts, forum entries, reviews) that contain topics configured by a WASM user.

[1] https://watson.analytics.ibmcloud.com

Proceedings of the EACL 2017 Software Demonstrations, Valencia, Spain, April 3-7 2017, pages 53–56

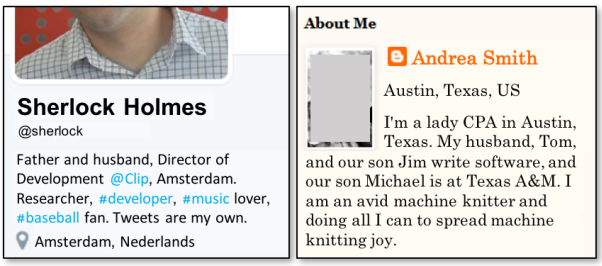

Figure 1: Example author biographies.

Our segmentation approach only relies on the 'on-topic' content of an author, not on all of the author's posts—for two reasons: Firstly, many social media sources such as forums do not allow retrieving content based on an author name. Secondly, social media content providers charge per document. It would not be economical to download all content for all authors who talk about a certain topic. We found that the combination of the author's content with the author's name(s) and biography significantly enhances both precision and recall of audience segmentation.

2.1 Author Location

WASM detects the *permanent* location of an author, i.e., the city or region where the author lives, by matching the information found in their biography against large, curated lists of city, region and country names. We use population information (preferring the location with higher population) to disambiguate city names in the absence of any region or country information, e.g., an author with the location *Stuttgart* is assigned to Stuttgart, Germany, not Stuttgart, AR, USA.

2.2 Author Demographics

WASM identifies three demographic categories:

Marital status: When the author is married, the value is *true*, otherwise*unknown*. For the classification we identify keywords (encoded in dictionaries) and patterns in the author's biography and content. Our dictionaries and patterns match, for example, authors describing themselves as *married* or *husband*, or authors that use phrases such as *my spouse loves running* or *at my wife's birthday*. Thus, WASM tags the authors in Figure 1 as married.

Parental status: When the author has children, the value is *true*, otherwise *unknown*. Similarly to the marital status classification, keywords and patterns are matched in the authors' biographies and contents. Example keywords are *father* or *mom*; example patterns are *my kids* or *our daughter*. WASM tags both authors in Figure 1 as having children.

Gender: Possible values are *male*, *female* and *unknown*. The classification of gender relies on a similar keyword and pattern matching as described previously. In addition, it relies on matching the author name against a built-in database of 150,000 first names that span a variety of cultures. For ambiguous first names such as *Andrea* (which identifies females in Germany and males in Italy), we take both the language of the author content and the detected author location into account to pick the appropriate gender. WASM classifies the first author in Figure 1 as male, the second author as female.

2.3 Author Behavior

WASM identifies three types of author behavior: **Users** own a certain product or use a particular service. They identify themselves through phrases such as *my new PhoneXY*, *I've got a PhoneXY* or *I use ServiceXY*. **Prospective users** are interested in buying a product or service. Example phrases are *I'm looking for a PhoneXY* or *I really need a PhoneXY*. **Churners** have either quit using a product or service, or have a high risk of leaving. They are identified by phrases such as *Bye Bye PhoneXY* or *I sold my PhoneXY*.

Author behavior is classified by matching keywords and patterns in the authors biographies and content—similarly to the demographic analysis described above. It allows to understand: *What do actual customers of a certain product or service talk about? What are the key factors why customers churn?*

Figure 2 shows an example analysis where the topics of interest were three retailers: a discounter, an organic market and a regular supermarket.[2] The top of Figure 2 summarizes what is relevant for authors that WASM identified as *user*s, i.e., customers of a specific retailer. The bottom part shows two social media posts from *user*ss.

2.4 Author Interests

WASM defines a three-tiered taxonomy of interests, which is inspired by the IAB Tech Lab Content Taxonomy (Interactive Advertising Bureau, 2016). The first tier represents eight top-level interest categories such as *art and entertainment*

[2]The retailers' names are anonymized.

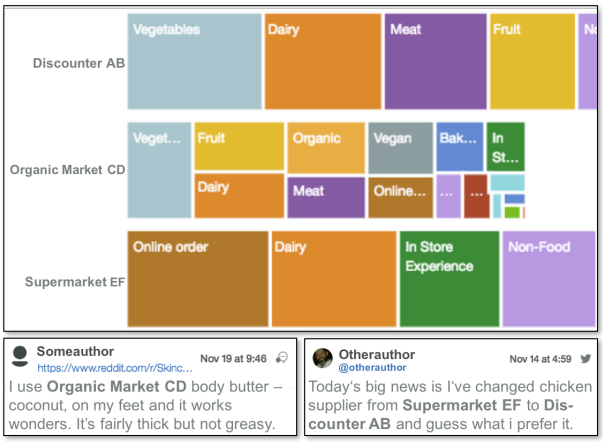

Figure 2: What matters to users of supermarkets.

or *sports*. The second tier comprises about 60 more specific categories. These include *music* and *movies* under *art and entertainment* and *ball sports* and *martial arts* under *sports*. The third level represents fine-grained interests, e.g., *tennis* and *soccer* below *ball sports*.

The fine-grained interests on the third level are identified in author biographies and content with the help of dictionaries and patterns. From the biography of the first author in Figure 1, we infer that he is interested in music and baseball (*music lover* and *baseball fan*). For the second author we infer an interest in machine knitting (*I am an avid machine knitter*). Note that a simple occurrence of a certain interest, e.g. a sport type, is usually not enough: it has to be qualified in the author content by matching specific patterns. A match in a biography, however, typically qualifies as a bona fide interest—excluding, e.g., negation structures.

Figure 3 visualizes the connections between the three retailers mentioned in the previous chapter and coarse-grained interest categories. This reveals that authors who write about *Organic Market CD* are uniquely interested in animals and that *Discounter AB* does not attract authors who are interested in sports. These insights allow targeted advertisements or co-selling opportunities.

Figure 3: Author interests related to Retailers

2.5 Account Type Classification

WASM classifies social media accounts into *organizational* or *personal*. This helps customers to understand who is really driving the social media conversations around their brands and products: is it themselves (through official accounts), a set of news sites, or is it an 'organic' conversation with lots of personal accounts involved?

Organizational accounts include companies, NGOs, newspapers, universities, and fan sites. Personal accounts are non-organizational accounts from individual authors. Account types are distinguished by cues from the authors' names and biographies. For example, company-indicating expressions such as *Inc* or *Corp* and patterns like *Official Twitter account of <some brand>* indicate that the account represents an organization, whereas a phrase like *Views are my own* points to an actual person. The biographies in Figure 1 contain many hints that both accounts are personal, e.g., agent nouns (*director*, *developer*, *lover*) and personal and possessive pronouns (*I*, *my*, *our*).

3 Implementation

3.1 Text Analysis Stack

The WASM author segmentation components are created by using the *Annotation Query Language* (AQL; Chiticariu et al., 2010). AQL is an SQL-style programming language for extracting structured data from text. Its underlying relational logic supports a clear definition of rules, dictionaries, regular expressions, and text patterns. AQL optimizes the rule execution to achieve higher analysis throughput.

The implemented rules harness linguistic preprocessing steps, which consist of tokenization, part-of-speech tagging and chunking (Abney, 1992; Manning and Schütze, 1999)—the latter also expressed in AQL. Rules in AQL support the combination of tokens, parts of speech, chunks, string constants, dictionaries, and regular expressions. Here is a simplified pattern for identifying whether an author has children:

```
create view Parent as
extract pattern
  <FP.match> <A.match>{0,1} /child|sons?|kids?/
from FirstPersonPossessive FP, Adjectives A;
```

It combines a dictionary of first person possessive pronouns with an optional adjective (identified by its part of speech) and a regular expression matching child words. The pattern matches

phrases such as *my son* and *our lovely kids*.

3.2 System Architecture

We wanted to provide users analysis results as quickly as possible. Hence, we created a *data streaming* architecture that retrieves documents from social media, and analyzes them on the fly, while the retrieval is still in progress. Moreover, we wanted to run all analyses for all users on a single, *multi-tenant* system to keep operating costs low. The data stream that our text analysis components see contains social media content for different users. Hence, we built components that can switch between different analysis processing configurations with minimal overhead.

The text analysis components run as Scala or Java modules within Apache Spark™. We exploit Spark's Streaming API for our data streaming requirements. The data between the components flows through Apache Kafka™. The combination of Spark and Kafka allows for processing scale-out, and resilience against component failures.

The multi-tenant text analysis processing is separated into user-specific and language-specific analysis steps. We optimized the user-specific analysis steps (such as detecting products and brands a particular user cares about) to have virtually no switching overhead. The language-specific steps (e.g., sentiment detection or author segmentation) are invariant across customers. To achieve the required processing throughput, we launch one language-specific rule set per processing node, and analyze multiple documents in parallel with this language-specific rule set.

The analysis pipeline runs on a cluster of 10 virtual machines, each with 16 cores and 32G RAM. Our customers run 300+ analyses per day, analyzing between 50,000 to 5,000,000 documents each.

4 Evaluation

Our customers are willing to accept smaller segment sizes (i.e., lower recall) as long as the precision of the segment identification is high. This design goal of our analysis components is reflected in the evaluation results for author behavior and account types on English social media documents that we present here.[3]

The retail dataset mentioned in previous chapters consists of 50,354 documents by 41,209 authors from all social media sources. WASM classifies 1,695 authors as *users*—without any additional effort by the customer. Compare that with the task to run a survey in the retailer's stores (and its competition) to get similar insights. We manually annotated author behavior for 500 documents from distinct authors. The evaluated precision is 90.0% at a recall of 58.3%.

The evaluation of account types is based on a random sample of 50,124 Tweets, which comprises 43,193 distinct authors. 36,682 of the authors provide biographies. We use this as the "upper recall bound" for our biography-based approach. Our system assigns an account type for 18,657 authors, which corresponds to a recall of 50.9%. It classifies 16,981 as *personal* and 1,676 as *organizational* accounts. We manually annotated 500 of the classified accounts, which results in a precision of 97.4%.

5 Conclusion and Future Work

This paper presented real-life use cases that require understanding audience segments in social media. We described how IBM Watson Analytics for Social Media™ segments social media authors according to their location, demographics, behavior, interests and account types. Our approach shows that the author biography in particular is a rich source of segment information. Furthermore, we show how author segments can be created at a large scale by combining natural language processing and the latest developments in data analytics platforms. In future work, we plan to extend this approach to additional author segments as well as additional languages.

[3] An evaluation of each audience feature for each supported language is beyond the scope of this paper.

References

Steven P. Abney, 1992. *Parsing By Chunks*, pages 257–278. Springer Netherlands, Dordrecht.

Laura Chiticariu, Rajasekar Krishnamurthy, Yunyao Li, Sriram Raghavan, Frederick Reiss, and Shivakumar Vaithyanathan. 2010. SystemT: An Algebraic Approach to Declarative Information Extraction. In *Proceedings of ACL*, pages 128–137.

Interactive Advertising Bureau. 2016. IAB Tech Lab Content Taxonomy. Online, accessed: 2016-12-23, https://www.iab.com/guidelines/iab-quality-assurance-guidelines-qag-taxonomy/.

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, USA.